Tampere University

Otto Sahlgren

ALGORITHMIC DECISION-MAKING, DISCRIMINATION AND DISRESPECT

An Ethical Inquiry

Faculty of Social Sciences
Master's Thesis
March 2020

ABSTRACT

Otto Sahlgren: Algorithmic Decision-making, Discrimination and Disrespect – An Ethical Inquiry
Master's Thesis
Tampere University
Degree Program in Philosophy
March 2020

The increasing use of algorithmic decision-making systems has raised significant legal and ethical concerns in several contexts of application, such as hiring, policing and sentencing. A range of literature in AI ethics shows how predictions and decisions generated on the basis of patterns in historic data may lead to discrimination against different demographic groups – those that are legally protected and/or are in positions of vulnerability, in particular. Both in the literature and public discourse, objectionable algorithmic discrimination is commonly identified as involving discriminatory intent, use of sensitive or inaccurate information data in decision-making, or as involving unintentional reproduction of systemic inequality. Some claim that algorithmic discrimination is inherently objectifying, unfair, and others find issue in the use of statistical evidence in high-stakes decision-making altogether. As is exemplified by this list of claims, the discourse exhibits considerable discrepancies regarding two questions: (i) how does discrimination arise in the development and use of algorithmic decision-making systems and (ii) what makes a given instance of algorithmic discrimination impermissible? Notably, the discussion around biased algorithms seems to have inherited conceptual problems that have long characterized the discussion on discrimination in legal and moral theory.

This study approaches the phenomenon of algorithmic discrimination from the point of view of ethics of discrimination. Through exploring Benjamin Eidelson's pluralistic, disrespect-based theory of discrimination, in particular, this study argues that while some instances may be wrong due to the issues with accuracy, unfairness, and algorithmic bias, the wrongness of algorithmic discrimination cannot be exhaustively explained by reference to these issues alone. This study suggests that some prevalent issues with discrimination in algorithmic decision-making can be traced to distinct choices and processes pertaining to the design, development and human-controlled use of algorithmic systems. However, as machine learning algorithms perform statistical discrimination by default, biased design choices and issues with "human-in-the-loop" enactment of algorithmic outputs cannot offer the full picture as to why algorithmic decision-making may have a morally objectionable disparate impact on different demographic groups.

Applying Eidelson's account – albeit with minor modifications – the wrongness of algorithmic discrimination can be explained by reference to the harm it produces, the demeaning social meaning it expresses, and the disrespectful social conduct it sustains and exacerbates by reinforcing stigma. Depending on context, algorithmic discrimination may produce significant individual and societal harms as well as reproduce patterns of behavior that go against the moral requirement that we treat each other both as moral equals and as autonomous individuals. The account also explains why formally similar but idiosyncratic instances of algorithmic discrimination which result in disadvantage for groups that are not specified by socially salient traits, such as gender, may not be morally objectionable. A possible problem with this account stems from lack of transparency in algorithmic decision-making: in constrained cases, algorithmic discrimination may be morally neutral if it is conducted in secret. While this conclusion is striking, the account is both more robust in comparison to alternative accounts, and defensible if one understands transparency as a pre-condition for the satisfaction of multiple other ethical principles, such as trust, accountability, and integrity.

The study contributes to the discussion on discrimination in data mining and algorithmic decision-making by providing insight into both how discrimination may take place in novel technological contexts and how we should evaluate the morality of algorithmic decision-making in terms of dignity, respect, and harm. While room is left for further study, the study serves to clarify the conceptual ground necessary for engaging in an adequate moral evaluation of instances of algorithmic discrimination.

Keywords: Algorithmic decision-making, artificial intelligence, discrimination, fairness, ethics, disrespect

CONTENTS

Introduction

Advances in computer and data science have made it possible to utilize extensive amounts of data in decision-making and to automate various practices that previously necessitated greater human involvement and effort. Decision-making processes are increasingly becoming partly or wholly delegated to so-called *algorithmic decision-making systems* (hereby 'ADSs'). These complex technological systems – also referred to as artificially intelligent systems (AI) – sort, classify, rank, and score individuals and events on the basis of statistical models that are extracted from data on previous human-made decisions. To do this, machine learning (ML) algorithms are used for "mining" the data; to find novel patterns in it, which may provide valuable insight to decision-makers. ML algorithms may also be used to train or teach an ADS by showing it examples of past decisions. By training it on extensive amounts of data, it will approximate a decision function from input data to a given output – that is, from data to a decision. Once trained, an ADS may be used either to advise and inform humans in decision-making or to autonomously execute decisions. (Barocas & Selbst 2016; Mittelstadt et al. 2016; Citron & Pasquale 2014; Zarsky 2014.) A movement towards algorithmic decision-making (AD) can be seen in both the private and public sector (Kelleher & Tierney 2018, 24–25). Prominently, ADSs are used to rank job candidates' resumes (cf. Dastin 2018), to assign credit scores (cf. Citron & Pasquale 2014), to predict hotspots for criminal activity and allocate police resources (cf. Selbst 2017), and to assess defendants' risk for recidivism (cf. Dressel & Farid 2018), to name just a few examples.

This development has not come without ethical and legal concerns, however. So-called algorithmic discrimination has become a paramount issue, as algorithmic decisions have been found to discriminate against individuals, those belonging historically disadvantaged groups, in particular. COMPAS[1], a criminal risk assessment tool used by U.S. courts to inform parole, pretrial and sentencing decisions, has been notoriously accused of being racially biased in the predictions it produces (Angwin et al. 2016); the use of predictive policing algorithms, such as PredPol, has led to over-policing in neighborhoods that are primarily inhabited by ethnic minorities (Lum & Isaac 2016; see also Selbst 2017); women are disadvantaged when algorithms are used to recommend suitable candidates for jobs (cf. Dastin 2018; see also West et al. 2019); and image and facial recognition systems systematically misclassify women – black women, in particular – at disproportionate rates in comparison to other groups (cf. Buolamwini & Gebru 2018).

[1] COMPAS is developed by Equinox (former Northpointe). The acronym stands for "Correctional Offender Management Profiling for Alternative Sanctions".

Several types of concerns related to algorithmic discrimination have been expressed in the public and scholarly discourse, as distinguished by Barocas (2014). Firstly, some worry that data mining may enable decision-makers to intentionally discriminate against individuals on the basis of their membership in legally protected groups. Data mining may provide decision-makers means for accurately inferring sensitive information about individuals (e.g. one's gender), which may then be used against them by bad actors in discriminatory intent. Call this the *bad actor view* (BAV). Secondly, some have expressed concerns about statistical bias and erroneous inferences. Poor sampling and use of unreliable statistical methods, many have argued, may result in unfairness – e.g. higher error rates in decisions for certain demographics – when the ADS is deployed. Thus, the concern seems to not be only bad actors but also unintentional discrimination and bad technology. Call this the *bad technology view* (BTV). A third concern is that even in the absence of significant bias or discriminatory intent, existing inequalities and systematic disadvantage suffered by vulnerable groups may be reproduced via AD: Disproportionalities in groups' baselines with respect to some attribute, such as arrest rates, are reflected in the training data which function as ground truth for future algorithmically generated predictions. The problem is that while in some cases these baseline differences may constitute justified bases for decision-making, in other cases they might in fact be explained by past discrimination and oppressive practices, such as racially biased policing. Call this the *structural discrimination view* (SDV). The possible adverse effects on legally protected groups (i.e., "disparate impact" or "indirect discrimination" in legal terms) are further aggravated by two issues. Firstly, identifying and preventing objectionable algorithmic discrimination is difficult due to lack of transparency in the design and use of ADSs, as well as the inherently opaque nature of these systems as such. Secondly, the biased predictions of an algorithm may become self-fulfilling; negative spirals may ensue when previous, negatively affecting predictions breed new ones by way of feeding back into subsequent decision-making processes in multiple contexts. (Barocas 2014; Citron & Pasquale 2014; Zarsky 2014.)

The worries concerning discrimination expressed in the discourse and literature differ significantly along several lines. First, they differ in their conceptions concerning *how discrimination may arise* through data mining and decision-making processes that follow it. In other words, they identify distinct "mechanisms" – (sequences of) acts and processes – that ought to be considered discriminatory. (Barocas 2014.) Secondly, they assume different positions on the question of what specifically makes a given instance of such discrimination *wrongful*. Some of these views take the accuracy of algorithmic decisions as a key issue; bad data leads to bad predictions. Others focus on the role of intention; AD offers malicious agents novel technological tools for engaging in discrimination. Yet ADSs may also discriminate on the basis of accurate inferences about individuals

that may, nevertheless, reflect past injustice or only map out symptoms of larger underlying social problems. Moreover, it seems that this may also be an unintended outcome of AD as algorithms are often opaque with respect to the processing behind their predictions. Lastly, a tension between distinct notions of fairness seems to underlie these views (Barocas 2014.) "Fair ML" has become a key issue in the literature on AI ethics, and an abundance of distinct metrics and definitions for fairness have been presented (cf. Verma & Rubin 2018). But what constitutes fairness in decision-making? Proponents of procedural fairness emphasize that individuals have equal claim to a fair procedure of decision-making; an algorithm ought to remain blind to sensitive traits of individuals, such as gender, for example. This may be an undesirable demand if taken categorically, however, as one might have to rely on sensitive information to correct past inequalities. Some understand fairness through equitable distributions in outcomes; different groups ought to receive roughly equal amounts of positive and negative decisions, or they should have equal misprediction rates. However, in some contexts requiring predictive equality may conflict with other reasonable and morally praiseworthy goals – "fair" algorithms may also come with a cost to society (cf. Corbett-Davies et al. 2017).

Acknowledging these discrepancies, Barocas concludes that "the current debate suffers from many of the same conceptual challenges that have characterized discrimination from its very inception as a formal notion in the law" (2014, 4). Indeed, legal and moral philosophers have long examined the perplexing questions concerning discrimination, equality, and fairness in the abstract. The difficulties that characterize these debates seem to have also translated into the context of algorithmic discrimination. Gaining clarity on the notion of (algorithmic) discrimination is crucial, however. Conceptual vagueness risks both overlooking instances of objectionable discrimination as well as shoehorning other moral issues, however significant they may be, as claims of discrimination.

In this study, I examine the notion of algorithmic discrimination by drawing on the philosophical literature on ethics of discrimination. Specifically, I consider how a theory that reduces the moral wrong of discrimination into the moral general wrong of *disrespect* could be applied in the context of AD. The central question is, then, whether a disrespect-based theory of discrimination can both (i) identify instances of wrongful algorithmic discrimination and (ii) explain why they are wrong. The first question is about identifying what types of instances of AD ought to be understood as wrongful discrimination. The second is about explaining why these instances are morally objectionable while others may be morally neutral or even praiseworthy. This also requires that one can distinguish objectionable instances of discrimination from other moral issues that may be wrong, but not inherently connected to discrimination. Two points of clarification are necessary at this point. Firstly, my analysis concerns the concept of discrimination in moral theory, rather than in jurisprudence, although there are relevant connections between the two. In other words, this study

examines the *ethics* of algorithmic discrimination, irrespective of any legal concept (or prohibition) of discrimination. Secondly, I limit my examination to the use of ADSs in high-stakes decisions, such as credit-scoring and recidivism risk assessment. This excludes considerations of so-called recommender systems that learn from users' preferences and enable personalization of online content, for example. Albeit similar in terms of the used technology, I focus on ADSs due to their increasing use in public governance and other contexts in which decisions have a clear and significant impact in individuals' social and economic status. This is not to undermine the significance of the possible effects – both good and bad – recommender systems may have, however.

The study proceeds as follows. In chapter 1, I examine the concept of discrimination in moral theory. As moral theorists have noted, the term 'discrimination' commonly bears a loaded meaning, but it ought to be understood as lending no *a priori* judgment regarding the moral status thereof. Thus, it is necessary to distinguish between a non-moralized sense of term and a moralized sense. The chapter provides a view of what discrimination consists *as an act* and offers an overview of different types of discrimination. These include direct, second-order, structural and statistical discrimination. I proceed to examine how the wrongness of discrimination could be explained in terms of disrespect. I provide a brief examination of different conceptions of disrespect, and I suggest that a view in which disrespect is understood as a deliberative failure to afford normative weight to the personhood of an individual fares better over two others. Such a view is presented by Benjamin Eidelson in *Discrimination and Disrespect* (2015). I proceed to examine, and slightly modify, Eidelson's pluralist theory of discrimination, which maintains that discrimination can be either intrinsically or contingently wrong. It will be intrinsically wrong if and when it manifests disrespect towards the personhood of those discriminated against (i.e., the discriminatees). In Eidelson's account, discrimination will be contingently wrong if and when so-called *broad harms* produced in the process overweigh its benefits. These include both individual material and psychological harms as well as social harms, such as corrosion of mutual respect between individuals and groups in communities. While I offer a preliminary overview of contingently wrongful discrimination in Eidelson's account, I will further elaborate his arguments in chapter 4.

In chapter 2, I provide an overview of the design and development process of ADSs. The overview comprises a general picture of how ADSs are designed, how they can be trained on historical data and ultimately used in decision-making. Special focus is on the role of subjective human deliberation in the design and use of ADSs. The overview will be accompanied by a mapping of risks and methodological pitfalls related to each stage of the design process, which may introduce so-called *algorithmic bias* into the model, possibly leading to impermissible discrimination as a result of AD. I conclude the chapter by examining the phenomenon of algorithmic discrimination in terms

of distinct acts – including different types of discrimination – taking place at distinct dimensions of conduct. I suggest that some of the discrepancies in the discourse on algorithmic discrimination are partly explained, on the one hand, by the fact that there are diverse mechanisms of discrimination in AD and, on the other hand, by the fact that the term "algorithmic bias" is used in many distinct senses. Specifically, I suggest that the outcomes of certain instances of algorithmic discrimination may be partly explained by preceding second-order and structural discrimination in the design process, in addition to statistical discrimination for which the ADS itself is being used. Thereby, some instances of algorithmic discrimination may comprise many "layered" instances of discrimination, some of which are not algorithmic in the sense that the relevant discriminatory acts would involve the use of algorithms. Furthermore, I claim that the moral evaluation of so-called "unalloyed" algorithmic discrimination – well-intentioned use of accurately performing ADSs that serves a reasonable goal (e.g., crime prevention) and which is unaccompanied by other contingent moral issues, but which has a disparate impact on a certain group – should be understood as concerning the question of whether and when it is impermissible to discriminate against individuals on the basis of non-universal generalizations (e.g., profiles, stereotypes and statistical evidence). This question should be considered separately from discrimination that occurs in the design process and other contingent issues, wrongness of which may be more straightforwardly explainable, at least in theory.

In chapter 3, I consider principled objections against statistical discrimination and, by extension, algorithmic discrimination. Specifically, I evaluate two types of arguments in defense of the notion that algorithmic discrimination is intrinsically wrongful as a form of discrimination when it has a disparate impact on vulnerable or legally protected groups. Allegedly, algorithmic discrimination will be disrespectful of persons' individuality and autonomy due to issues related to objectification, inaccuracy or the probabilistic nature of the evidence used to justify differential treatment of individuals. Alternatively, algorithmic discrimination will be inherently unfair by not meeting the requirement of impartiality and thus fails to treat people as equals. I refute these claims by demonstrating that algorithmic discrimination does not necessarily involve disrespect for individuality or unfairness understood as unequal treatment. I also argue that statistical evidence may constitute an epistemically justified basis for differential treatment, given that it is reasonably accurate. I conclude that unalloyed instances of algorithmic discrimination should not be considered wrongful irrespective of social contingencies, such as the social meaning they express.

In chapter 4, I consider Eidelson's Broad Harms Argument as a way of explaining how and when algorithmic discrimination may be considered contingently wrong. According to this view, one should deem even unalloyed algorithmic discrimination wrongful if and when it causes material, psychological and broad societal harms that overweigh the benefits of implementing the technology.

The prominent harms may be expressive; they are produced in virtue of the relation between a given practice and so-called "background injustice", such as injustice and oppression that demographic groups suffer and have suffered from historically. Notably, this account leads to a dilemma in constrained contexts: If we assume that algorithmic profiling may produce significant benefits and that these benefits are not outweighed by material harms resulting from profiling, unalloyed algorithmic discrimination may be morally neutral given that it is conducted in secret. I argue that while this notion is striking, it is defensible. To conclude the study on a hopeful note, I present a brief overview of some ways in which wrongful algorithmic discrimination can be tackled.

This study contributes to the discussion on discrimination in data mining and AD by providing insight into both how discrimination may take place in novel technological contexts and how we should evaluate the morality of algorithmic decision-making in terms of dignity, respect, and harm. While room is left for further study, the study serves to clarify the conceptual ground necessary for engaging in an adequate moral evaluation of instances of algorithmic discrimination.

1. What is Discrimination and When is it Wrong?

The concept of discrimination holds several distinct yet overlapping meanings in common use. One can discriminate between colors, in one sense of the term, by perceiving relevant differences between them. Another sense concerns differential treatment of individuals and groups, and the legal basis or justification thereof. This concept of discrimination is reflected in anti-discrimination legislation which provides lists of traits on the basis of which differential treatment is legally prohibited (i.e., legally protected groups), and hints at the contexts and sites which are of special concern in this regard (e.g. education or employment).[2] In this study, what is examined is a third sense of the term: the concept of discrimination in moral philosophy and ethics.

First, a distinction is to be made between a descriptive, non-moralized concept of discrimination and a moralized one (cf. Altman 2016, Sec. 1.2.; Lippert-Rasmussen 2014, 1). On the one hand, the non-moralized concept of discrimination refers to certain kinds of differential treatment of individuals or groups, without reference to the moral permissibility of the act. A definition for discrimination in this sense specifies the conditions that an act must satisfy for it to fall under the category of discriminatory acts. That is, any descriptive account of discrimination should explain the difference between discriminatory acts and other types of acts, such as non-discriminatory differential treatment or lying[3]. The moralized notion of discrimination, on the other hand, involves the idea that discrimination (as an act) may be considered at least *prima facie* objectionable.[4] (Altman 2016, Sec. 1.2.) An account of what makes discrimination morally objectionable is, then, a theory stating what other conditions need to be met for a given instance of discrimination to be morally wrong. As Erin Beeghly states, a robust theory of discrimination should do two things: (1) "identify wrongful cases of discrimination as such" and (2) "adequately explain why wrongful cases of discrimination are wrong" (2017, 87). The former condition may be called *the identification condition* and the latter *the explanation condition* (Ibid.). The central question of this study, then, is whether and how a disrespect-based theory of discrimination will distinguish genuinely wrongful instances of

[2] For example, Article 21 in the *European Union's Charter of Fundamental Rights* lists "sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation" as protected attributes. Article 23 states that "[e]quality between women and men must be ensured in all areas, including employment, work and pay." However, this is a prohibition of discrimination only in the public sphere. Inequalities of the private sphere (e.g. maintaining strictly gendered roles in households or exclusively preferring individuals of certain ethnicities as romantic partners) remain outside this scope.

[3] For example, a hiring process may involve differential treatment when two candidates are hired for different yet equally desirable jobs, but this seems different from a hiring decision which favors one candidate over another

[4] As Altman notes, it is common that the term 'discrimination' is used in a moralized sense to denote a certain category of morally objectionable acts (2016, Sec. 1.2). It is important to distinguish the non-moralized and moralized concepts, however, because there is disagreement among theorists as to whether all instances of discrimination should be considered morally objectionable.

algorithmic discrimination from ones that are not wrong, and ones that are confounded with other moral issues. Furthermore, it should explain what makes those instances wrongful.

Theories of discrimination differ in whether instances of wrongful discrimination are considered *intrinsically* or *contingently* wrong. Discrimination of the former kind is wrong regardless of its consequences or knowledge of those consequences. Conversely, discrimination that is contingently wrongful is wrong not because of anything intrinsic to the act itself but, rather, in virtue of its outcomes or other contingencies related to the act.[5] Harm-based accounts of wrongful discrimination (cf. Lippert-Rasmussen 2014), for instance, take discrimination to be always contingently wrong – if at all. Reflecting a consequentialist approach to ethics, harm-based accounts posit that the only way to evaluate whether some discriminatory act is wrong is to evaluate the harms and benefits that result from it. Similarly, Hellman's (2008; 2017) account posits that discrimination is wrong if (and when) it expresses that members of some group are morally inferior than members of another. By contrast, an account according to which the outcomes of discrimination (or other contingent features) bear no significance for its moral objectionability would deem all instances of wrongful discrimination intrinsically wrong. Finally, pluralist accounts (cf. Alexander 1992; Eidelson 2015) hold that a discriminatory act may be either intrinsically or contingently wrongful, if at all[6].

In this study, I will focus on Benjamin Eidelson's pluralist theory and the question of how it could be applied into the context of AD. I examine how the theory can be used to dissect the notion of algorithmic discrimination and evaluate the morality of different types of instances of it, which exemplify different types of moral issues and types of acts.

1.1. Generic Features and Types of Discrimination

I will first consider how one might define discrimination in the non-moralized sense. I start by examining one prominent definition, namely, that presented by Benjamin Eidelson in *Discrimination and Disrespect* (2015). I will also consider alongside some features Kasper Lippert-Rasmussen (2006; 2014) attributes to discriminatory acts (in the non-moralized sense) as their accounts vary only slightly in their nuances. After this, I provide an overview of different types of discrimination. These include *direct discrimination*, *second-order discrimination*, *structural discrimination*, and *statistical discrimination*. Each one is characterized by distinct features attributed to certain types of discriminatory acts and policies, with the exception of structural discrimination which does not

[5] That some act is contingently wrong does not mean it would be *less* wrong than an intrinsically wrongful act, however.
[6] Notably, Eidelson holds that all *acts* are contingently wrong, but he maintains that proper objects of moral evaluation are acts as they are "thickly" described; one's intentions are relevant for whether a certain act is wrongful or not. In this sense, the way an agent *comes to act* on the basis of her intentions, beliefs or desires may be intrinsically wrongful albeit the relevant act in itself ("thinly" described) may be contingently wrongful. This notion will be examined in chapter 1.2.4.

constitute genuine discrimination, according to Eidelson (2015, 25). I will first examine the concept of direct discrimination and elaborate on some generic features of discriminatory acts.

1.1.1. *Direct Discrimination*

Established in the late 19th century in the U.S., the Jim Crow laws functioned as an explicit means to enforce racial segregation. An enactment of these laws, a public transport policy in Montgomery, mandated that people of color sit at the back end of buses – a policy which ultimately became the focal point of a civil protest following Rosa Parks' arrest. This malicious policy is an instance of a directly discriminatory rule: when the policy is acted on, an individual is both intentionally and explicitly treated differently because she belongs to a particular group or has a certain trait[7]. Indeed, direct discrimination is often characterized by discriminatory intent and explicitness; it involves conscious reliance on "representational items – e.g. desires, beliefs, statements, laws – that refer to, or otherwise distinguish between, those who are discriminated against and those who are not in the relevant discriminatory respect" (Lippert-Rasmussen 2006, 170; see also Altman 2016, Sec. 2).

But consider a hiring process, where both a recruiter and an applicant are unaware that the recruiter is (*de facto*) making the decision not to hire the applicant based on the applicant's age. Such a case would not meet the requirements of intentionality or explicitness. It seems that it would be "direct" all the same, however; the decision is explained by the recruiter's unconscious regard for the applicant's age. Indeed, the aforementioned characteristics seem insufficient for identifying implicitly biased discrimination. A more nuanced definition is offered by Eidelson (2015, 17):

> X (directly) discriminates against Y in dimension W on the basis of P if and only if:
>
> | | |
> |---|---|
> | *(Differential Treatment Condition)* | X treats Y less favorably in respect of W than X treats some actual or counterfactual other, Z, in respect of W; and |
> | *(Explanatory Condition)* | a difference in how X regards Y P-wise and how X regards or would regard Z P-wise figures in the explanation of this differential treatment. |

The definition, presenting two necessary and jointly sufficient conditions for (direct) discrimination, seems to capture prominent instances of discrimination. However, it includes several notions that need be elaborated. I will first examine the variables that are at play in the definition.

[7] Paradigmatic examples of direct discrimination that do not involve policies or rules as such may often involve an explicit, even malicious, intention to discriminate against a certain group. A clear example would be where X discriminates against Y on the basis of a racist *belief* – i.e., the racist belief would here explain X's treatment of Y.

'X' denotes the agent that treats the two individuals or groups denoted by the patient-variables 'Y' and 'Z' differently. For Eidelson both the agent and patient variables may range over "people, corporations, institutions, or groups of these" (2015, 17). Importantly, X, Y and Z range over agents and patients only; not acts or policies. (Ibid.) The Montgomery bus policy, for instance, may only discriminate against people of color in the sense of being *enacted* by agents, and should thus not be confused as a subject or an agent who is conducting the discriminatory treatment.[8] In other words, rules and policies themselves do not act, but they may be formed in discriminatory intent or enforced to that aim. "Dimension W", then, refers to "a set of exclusive options regarding how one treats someone"[9] (Ibid., 18). These may be options regarding whether one shakes hands with someone (or not) or hires them for a job (or not), for example. The reason for using the term 'dimension' is that these options are mutually exclusive with respect to each Y and Z; one "cannot both hire and not hire" the same individual, at least at a given time (Ibid.). This entails that discrimination should be understood as *treatment-specific*: it is something that can be ascribed to a single act or decision (Lippert-Rasmussen 2014, 20–22). Indeed, analytically inclined philosophers generally prefer to talk about *instances* of discrimination – i.e., local acts of differential treatment rather than global disadvantage faced by some groups. A discriminatory policy, for example, may be enacted multiple times, but each enactment constitutes a distinct act of discrimination, respectively. This conceptualization of discrimination does not imply that some group would not be disadvantaged simultaneously in multiple areas of life (see chapter 1.1.3 below). Rather, it is to say that for sake of conceptual clarity, discrimination should be considered something that occurs always in relation to some procedure or decision individually, such as a hiring decision (Ibid.).

Regarding the patient variables Y and Z, it should be explicated that the individual or group – "a set of particulars" – should not be understood here necessarily as the set that is specified by a given value of P, when the discriminatory treatment is explained by differences with respect to P (Eidelson 2015, 18). In other words, the (perceived) trait that X attends to is not necessarily the trait that is possessed by (members of) Y, although it could be so. This means two things: Firstly, P-wise discrimination against Y is not equivalent to discrimination against Y *qua* Y. Rather, it is

[8] Lippert-Rasmussen holds similarly that the agent variable may range over individuals (or natural persons), groups, organizations, companies and other legal persons, and social structures. He takes that the patient variable may also range over non-human animals. He has also argued that discrimination resulting from social structures or conduct by legal persons can be reduced to the actions taken, or regularities and patterns of behavior sustained by some set of individuals. Accordingly, in his definition of generic discrimination the agent variable is separated from the act or treatment, "Φ-ing", conducted by the agent. (Lippert-Rasmussen 2006, 168; 2014, 15–20.) As I understand, this serves to employ a distinction similar to Eidelson's.

[9] Although absent in his later formulation of the definition for generic discrimination in *Born Free and Equal* (2014), an earlier formulation in Lippert-Rasmussen (2006) included a similar notion of a "dimension *W*" denoting the relevant process, procedure, distribution of goods etc., with respect to which individuals are treated disproportionately.

discrimination against Y on grounds that Y is perceived to differ from Z P-wise. Elaborating on this point, Eidelson notes that "[m]any immigrants are discriminated against […] and in some contexts it is important both that it is immigrants who are discriminated against and that it is on the basis of race that they are discriminated against" (Ibid.). This is only to say that the discriminatory act is not necessarily explained by differences in the same trait that specifies the group that is discriminated against.[10] Secondly, it is X's perception of Y's possessing a trait (and Z's not possessing that trait) that explains the differential treatment. Lippert-Rasmussen notes that one should understand discrimination as *actual-properties independent*; the actual properties Y and Z have need not correspond to the properties they are believed by X to have (2014, 20). The belief, rather, explains why X treats Y the way he does. Eidelson, in turn, notes that belief is not necessary. What matters is how the agents *regards* the patients of the treatment P-wise, or "which of [X's] perceptions actually contribute to explaining the way he acts" (Eidelson 2015, 20). Implicit, unconscious cognitive biases towards certain groups of people are problematic to analyze as belief-states, he argues, and this is why he prefers to formulate the Explanatory Condition in a way that simply requires "P-perceptions" to play a role in explaining X's behavior; not that X is aware of them. (Ibid., 21–22.) Specifically, there are two distinct claims here: (i) X will discriminate against Y in cases where X is neither consciously aware of perceiving P-wise differences nor of the fact that these differences figure into the explanation of X's behavior, and (ii) X will discriminate against Y even if he is aware of perceiving these differences, but not aware of them playing a part in the explanation of his behavior. (Ibid.)

Importantly, the Differential Treatment Condition states that for some act to be considered discriminatory, it needs to be *disadvantageous* for Y, and not merely treat Y and Z differently. This is a central point about the disadvantage aspect inherent to discrimination. Talking about racial discrimination, Altman points out that mere "[d]ifferential treatment is symmetrical: if blacks are treated differently from whites, then whites must be treated differently from blacks" (2016, Sec. 1.1). But this does not seem to capture the notion of that discrimination always involves *favoring* someone over others. Discrimination should thereby be taken to always involve disadvantage for someone. This does not however, necessitate that this someone is worse off as a result of the agent's treatment *overall*. "An act can both be discriminatory and, simultaneously, confer an absolute benefit on those discriminated against, because the conferral of the benefit might be combined with

[10] Another example would be an employer discriminating against foreigners on the basis of names in job applications. Assume that the employer X's regard of a foreign applicant Y name-wise in some manner figures to the explanation of X's decision (not) to hire Y. Assuming that X has no direct way of assessing whether Y indeed is a foreigner, there are two possible ways this might occur: either X discriminates against Y on the basis of Y's name only, or X discriminates against Y on the basis of *both* Y's name and foreign background by inferring from Y's name that Y is a foreigner. In the latter case, the name is used as a proxy. (Eidelson 2015, 19–20.)

conferring a greater benefit on the members of the appropriate comparison group", Altman points out (Ibid.). The notion of disadvantage inherent to discrimination entails two things: First, discrimination has a built-in requirement for contrast. It is, in Lippert-Rasmussen's terms, "*essentially comparative with respect to individuals*" (2014, 16). That Y is disfavored requires that there is some Z that functions as the relevant contrasted individual or group for evaluating to how Y is treated. However, the notion of "counterfactual others" allows that such a group need not actually exist in order to function as a contrast-patient (Eidelson 2015, 17)[11]. Second, this "relevant disadvantage is interpersonal, not intrapersonal", although discrimination could "involve treating the same individuals differently over time"[12] (Lippert-Rasmussen 2014, 16). Arguably, discrimination may also be *reflexive*[13] in that one may well possess some trait and yet discriminate against others – or even oneself – possessing that same trait (Ibid., 21).

I will next move on to consider another type of discrimination, namely, second-order discrimination. Notably, the notions related to discrimination examined here (e.g. treatment-specificity, actual properties -independence, comparative disadvantage, and so on) may be taken to apply also to other types of discrimination examined below.

1.1.2. *Second-Order Discrimination*

Direct discrimination is often distinguished from so-called indirect discrimination (or disparate impact, in legal terms). Two distinctive differences between the two can be explicated: Firstly, the latter "is *structurally comparative* in the sense that it involves the disadvantaging of certain *groups* of persons in relation to others" (Khaitan 2017, 32; italics added). It concerns groups rather than individuals as such. Secondly, indirect discrimination commonly lacks explicit reference to the groups that are discriminated against. That is, a rule or policy may adversely affect group Y (in relation to group Z) even though it involves no explicit reference to Y or Z, respectively. To borrow Eidelson's example (2015, 28), a manager of a factory might require that potential employees exceed a certain threshold in upper-body strength which might result in women being disfavored in the process. That women are disadvantaged is here explained by baseline differences in upper-body strength between

[11] Lippert-Rasmussen holds this as well. Even possible people (or other possible objects), he argues, can function as contrast-patients. As an example, he presents a hypothetical situation where a person treats humans worse in comparison to a (possible or counterfactual) superior race of humans, which he believes to exist. (Lippert-Rasmussen 2014, 19.)

[12] One could, for example, discriminate against a former acquaintance on the basis of a recent shift in his views on politics, refusing to thereby acquaint oneself with him.

[13] To elaborate, Lippert-Rasmussen presents a hypothetical example of the latter scenario, where a woman that is better qualified for a job recommends a male co-worker for promotion solely on the basis of his gender, omitting for recommending herself. Similarly, the woman might think that a female co-worker is better qualified (than the man) and yet recommend the man for the promotion on the basis of his gender. Although Eidelson does not address intrapersonal and reflexive discrimination, his definition seems to account for such cases as well.

women and men. Of course, indirectly discriminatory policies may also track (or be "parasitic on") existing inequalities produced by past direct discrimination (Khaitan 2017, 38). For example, if promotion decisions are made on the basis of years worked at a firm, past discrimination of women or minorities in hiring may lead to the exclusion of members of these groups from the candidate pool. Notably, while lack of explicitness seems integral here, intention (or lack of it) does not seem to ground a distinction between direct and indirect discrimination. A policy can be enacted either *in order* to disadvantage Y or the disadvantage may be a mere unintended consequence of that policy. Intentionality may prove significant in evaluating the justification of both, yet it does not offer conclusive judgment as to whether indirect discrimination is objectionable.[14] (Altman 2016, Sec. 2.2.; Khaitan 2017.)

A distinctive feature of Eidelson's account, however, is that 'indirect' discrimination is not taken to constitute a genuine type of discrimination at all. He argues that the notion of 'indirect' discrimination involves an unwarranted stipulation that the *moral* similarity of direct and 'indirect' discrimination would constitute a similarity in the types of *acts* in question[15] (2015, 58). In other words, while instances of both would seem to bear some similarity with respect to why they are wrong, they should not thereby be regarded as distinct "species of discrimination", partly because "discrimination is at least not *only* a moral category" but a category of acts (Ibid., 52). Furthermore, if direct and 'indirect' discrimination are wrong, their wrongness might stem from different grounds. "[I]mposing continuity" on the two concepts "will therefore involve significant distortion" that one could avoid by conceiving indirect discrimination in a different manner (Ibid., 56). But how?

As noted, some rules or policies may be deliberately implemented in order to favor or disadvantage a group 'indirectly', e.g., to avoid accountability. In such cases, Eidelson argues, one would be better off talking about instances of so-called *second-order discrimination*, which are ultimately instances of direct discrimination. (2015; 41–44, 58.) Second-order discrimination is discrimination "on the basis of P in adopting a rule or decision to discriminate on the basis of Q in some other dimension of treatment" (Ibid., 41). In this sense, an instance of second-order discrimination temporally precedes an act (or forming of a rule or policy) that is commonly taken to exemplify so-called 'indirect' discrimination. This notion also accounts for those instances where this

[14] It may be that upper-body strength is a necessary requirement for working in a factory, for example. Moreover, simultaneously avoiding direct and indirect discrimination might prove difficult a task due to differences in groups' baselines with respect to some attribute. As Khaitan points out, "if one is to avoid inflicting disparate impact, one sometimes needs to intentionally use a protected characteristic, i.e., commit disparate treatment" (2017, 36). In fact, legal doctrines concerning disparate treatment (i.e. direct discrimination) and disparate impact (i.e. indirect discrimination) may sometimes be conflicted (Barocas & Selbst 2016, 725).

[15] Iris Young holds a similar view, according to which discrimination should be conceived only as an act-type that exemplifies intentionality and explicitness. For Young, indirect discrimination should, however, be understood in terms of oppression. (Young 1990, 196; Altman 2016, 3.1.)

sort of discrimination is unconscious yet stems from bias towards some group (although such cases might be rare). Essentially, Eidelson's claim is that discriminating on the basis of Q-wise differences – although they may map "the possible values of the trait" P – can be understood as an act of *direct* discrimination on the basis of differences with respect to Q (2015, 44).

To elaborate on this notion, assume that the factory manager's malicious intent is to discriminate on the basis of sex. Viewed as 'indirect' discrimination, the criterion for admission set by the manager would be considered discrimination on the basis of sex (P), despite the fact that the biased distribution in the outcomes of the hiring process is in fact explained by baseline differences in applicants' upper-body strength (Q). Eidelson contends that it seems more appropriate to say that the manager's regard of the candidates sex-wise figures into the explanation of *why he employs some criterion*, and that the set of people disfavored by that criterion (i.e., women) will be discriminated against in another *dimension* than hiring (W_2) – namely, in the dimension where that criterion is set or adopted (W_1). However, the set of people disfavored in virtue of adopting that criterion – those not exceeding a certain threshold of upper-body strength – will be discriminated in the dimension of hiring. An interesting implication of this internalist notion is that the men who do not exceed the threshold will also be "indirectly" discriminated against *on the basis of sex* just as much women, even though the manager aims to favor men. In a sense then, second-order discrimination may result in "false positives" and "false negatives" with respect to the discriminator's aims[16] because some men may not meet the criterion, while some women may[17]. (Eidelson 2015, 40–44.)

The concept of second-order discrimination serves to identify those instances of discrimination in which the discriminator (un)consciously discriminates against group Y in relation to Z on the basis of differences between them P-wise, but this discriminatory act takes place in a dimension that precedes some other dimension of treatment. However, if the agent's regard of the subjects sex-wise does not figure into the explanation of the treatment, "indirectly" discriminatory acts or policies should – according to Eidelson – not be confused with discrimination *per se*. Rather, they ought to be considered in terms of structural discrimination.

1.1.3. *Structural Discrimination*

The term 'structural discrimination' is often used to refer to the notion that certain groups are persistently disadvantaged by societal rules, norms, and institutions, in many ways and on multiple

[16] As discrimination is understood as actual-properties independent, such "false positives" and "false negatives" may follow from other types of discrimination as well, including direct discrimination. This may be, for example, if the discriminator mistakenly perceives Y to differ from Z P-wise.

[17] Here it is supposed that the set of people disfavored in the dimension of setting the criterion is not identical to those disfavored in the dimension of hiring – i.e., that not only women are disadvantaged by enforcing the hiring policy.

fronts. The conceptual focus is not on the intent or explicitness of discriminatory actions, however, but rather on the wide-ranging effects they have on historically subordinated groups, and the effect of maintaining that subordination (Pincus 1996, 190–191). Systemic disadvantage may be indirectly (re)produced by major institutions, rules and implicit norms and attitudes that govern and regulate peoples' social lives. However, empirically speaking it is most often the result of a long history of both directly and indirectly discriminating acts and policies. (Altman 2016, Sec. 2.3) Call this the *global* sense of the term 'structural discrimination'.

Eidelson's way of conceptualizing structural discrimination differs slightly from the global sense. As Eidelson and Lippert-Rasmussen argued (see chapter 1.1.1), it seems appropriate to understand discrimination in terms of distinct acts taking place in distinct dimensions. This in mind, Eidelson approaches structural discrimination as a question about mediation – or, rather, lack thereof:

> The essential difference between ordinary [i.e., direct] and structural discrimination […] is that in the latter case the explanatory connection between the comparative disadvantage and the trait in question need not be *mediated* by anything to do with how the (structural) discriminator *perceives or reacts to the trait* (Eidelson 2015, 25; italics added).

Again, Eidelson's focus is on acts as opposed to the outcomes or effects of discrimination as such. In his account, structural discrimination consists in acts (e.g. enactments of rules) which are seemingly neutral, but which burden some groups disproportionately "only because of underlying physical or social dynamics" which are connected to the traits that specify the disadvantaged groups (2015, 25). Notably, that this disproportionate burden is imposed on some groups may be explained by it tracking previous instances of direct, including second-order, discrimination (recall the example regarding promotion decisions). This is not necessary, however, as Eidelson points out:

> If an agency offers a vaccine which, unbeknownst to it, only works for men, it may not engage in any ordinary [i.e., direct] discrimination at all. But this act arguably constitutes structural sex discrimination against women: it is comparatively disadvantageous to them, and its being so is traceable to their being women. (2015, 25.)

Structural discrimination, thus, differs from direct and second-order discrimination in that differences between individuals P-wise do not explain why X acts in a way that burdens one group but not another. That is, the adverse effects on Y are not explained by how the discriminator perceives or values Y P-wise – the act does not meet the Explanatory Condition (with respect to P). The effects are, however, partly explained by discriminatory act, rule or policy tracking some baseline differences P-wise (be they social or physical). In this sense, whether the factory manager in our example is a

sexist second-order discriminator or a structural discriminator depends on whether or not imposing the upper-body strength criterion is explained by his (conscious or unconscious) regard of the candidates sex-wise.

Notably, Eidelson emphasizes that structural discrimination, in the relevant sense, does not actually constitute a sub-category of discriminatory acts due to this lack of mediation (2015, 26).[18] (He sees practical value in use of the term 'structural discrimination', nevertheless, and I will use it throughout this study, respectively.) Consequently, Eidelson's conceptualization also entails that prohibitions of 'indirect' and 'structural' discrimination not only aim to prevent wrongful second-order discrimination, but simultaneously serve a function of redistributive justice:

> [L]aws prohibiting [indirect discrimination] can best be understood as redistributive programs [that aim] to equalize opportunity for members of systematically disadvantaged groups. Far from a basic concept to be accommodated in a general account of the ethics of discrimination, then, indirect discrimination is in effect a construct for improving the status and welfare of disadvantaged groups indirectly by improving the representativeness of workforces in general. (Eidelson 2015, 67–68.)

In essence, Eidelson finds the legal concept of indirect discrimination useful "as a piece of legal jargon" but not as a robust concept of moral theory (2015, 19). Naturally, none of this entails that "indirectly" or structurally discriminatory acts and policies are morally unproblematic. Nor does it mean redistributive justice were not desirable. Rather, it highlights the notion that in light of the "divergent conceptual structures" of direct and 'indirect' discrimination, one ought to look for the reason for the moral objectionability of some 'indirectly' discriminatory acts and policies outside the scope of discrimination (Ibid., 58).

1.1.4. *Statistical Discrimination*

Belonging to a high-risk group for traffic accidents, young males typically have more costly insurance premiums. This method of differentiating insurance premium costs is an instance of *statistical discrimination*. (Lippert-Rasmussen 2007.) Statistical discrimination is instrumental in that it is "based on the belief, assumption, or fact, that" differentiating treatment on the basis of a *predictor trait* P will aid in achieving some "discrimination-independent goal" that involves finding some *target trait* T (Schauer 2017, 43). In our example, the higher cost is explained by there being evidence to

[18] Eidelson states he does not "mean to imply that it [structural discrimination] really is a kind of *discrimination*" (2015, 26). He says that "[p]refixed concepts are often ambiguous" because "they can either mark out a subspecies of a phenomenon, or they can qualify the claim that the thing in question falls under the concept at all", and that his definition "is of the latter kind" (Ibid).

the fact that 'being a young male' (P) is taken to be predictive of 'being involved in a traffic accident in the future' (T). In this sense, statistical discrimination is not explained by beliefs about fundamental differences between classes or groups of people, such as views about gender roles in society. The (value of a) predictor trait P only functions as a statistically evinced proxy for (the value of) some target trait T, thereby explaining why we should consider differences with respect to P when making decisions. Statistical generalizations are (most often), however, non-universal generalizations – i.e., akin to so-called *generic statements* (see chapter 1.3 below), they lack a universal quantifier[19]. Consequently, relying on statistical evidence in decision-making will most likely result in some number of false positives and false negatives, i.e., mispredictions and misfires.

A definition for statistical discrimination can be formulated by modifying Eidelson's definition of generic (direct) discrimination. I do this by supplementing the Differential Treatment Condition with what I call the Statistical Evidence Condition and the Differential Regard Condition, and by replacing the Explanatory Condition with the Joint Explanation Condition, respectively. Thus, we get the following definition for statistical discrimination:

X engages in statistical discrimination against Y in dimension W on the basis of P if and only if:

(*Differential Treatment Condition*)	X treats Y less favorably in respect of W than X treats some actual or counterfactual other, Z, in respect of W; and
(*Statistical Evidence Condition*)	Statistical evidence, E_S, accessed by X, suggests that differences P-wise predict or indicate differences T-wise; and
(*Differential Regard Condition*)	Y and Z are regarded by X to differ P-wise; and
(*Joint Explanation Condition*)	The Statistical Evidence Condition and Differential Regard Condition together – but not separately – explain why the Differential Treatment Condition is satisfied.

Notably, E_S may turn out to be spurious (or not). Whether or not there actually is a connection between a predictor and a target attribute, discrimination on the basis of statistical information about their connection will fall under statistical discrimination as an act-type. (Schauer 2017, 44–46.) Indeed, the strength of the link between (values of) P and T might vary significantly from case to case. As Schauer notes, "[t]here are strong and weak correlations, and thus strong and weak indicators, and thus indicators that are more or less reliable, or more or less accurate" (Ibid., 46). However, in the

[19] Of course, as a 100 % probability is still a probability, statistical evidence can in a certain sense comprise a universal generalization as well (regardless of whether the generalization is true).

literature on ethics of discrimination, it is often noted that the accuracy of non-universal generalizations (e.g. statistical evidence) does not explain the wrongness of acts that fall under the category of statistical discrimination (cf. Beeghly 2018, 689–691; Lippert-Rasmussen 2007, 390–391). This seems to be case with algorithmic discrimination as well, and I will argue why this is the case in chapter 3.

1.2. Disrespect-Based Theories of Discrimination

Having examined the non-moralized concept of discrimination and distinct types of generic discrimination, I will next consider the moralized notion of discrimination. As noted above, there are several moral theories of what grounds the moral objectionability of discrimination[20]. Disrespect-based theories – which I focus on here – hold that the moral objectionability of discrimination stems from a violation of a more fundamental moral principle: it manifests disrespect towards those who are discriminated against (i.e., the discriminatees). Beeghly (2017) distinguishes three conceptions of disrespect and, thus, three distinct disrespect-based theories of wrongful discrimination: the mental state theory (MST), the expressive theory (ET), and the deliberative failure theory (DFT). I will briefly consider each and suggest that the DFT offers the most robust account of disrespect among the three. Some aspects of the ET will be considered also later in the study, in chapter 4.3. Before I examine these views, however, it is necessary to consider the notion of *social salience*.

1.2.1. *Social Salience*

Commonly, (wrongful) discrimination is talked about in terms of specific social groups. Most anti-discrimination laws, for example, prohibit discrimination on the basis of an individual's religious beliefs, race, gender, and so on[21]. Aligning his harm-based account with this common notion, Lippert-Rasmussen introduces the social salience as a way of limiting the scope of wrongful discrimination to instances of discrimination that involve treating people differently on the basis of socially salient traits (cf. Lippert-Rasmussen 2006; 2014). For him, all instances of discrimination which are wrong (if they are) involve discrimination against socially salient groups. A group is understood as socially salient if "perceived membership of it is important to the structure of social interactions across a wide

[20] Several accounts of the wrongness of discrimination have been proposed in the literature, in addition to ones examined in this study. Irrelevance accounts posit that socially salient or sensitive categories are irrelevant for decision-making (cf. Halldenius 2017); desert-based accounts maintain that wrongful discrimination is explained by the fact that people are not treated according to what they deserve (cf. Moles 2017); some accounts hold that wrongful discrimination involves violations of personal freedom (cf. Moreau 2010; 2017) or unvirtuous conduct (cf. Garcia 2017). For an examination of different views see Altman 2016.

[21] Article 21 in the *European Union's Charter of Fundamental Rights* states that "[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."

range of social contexts" (Lippert-Rasmussen 2014, 30). Social salience is, thus, an empirical concept; it is *perceived* significance of some trait in different social contexts, and interactions between people and institutions. Gender, ethnicity, and religious beliefs, for instance, would arguably form the basis for a socially salient group in this sense, as they can be considered traits that play a significant role in individuals' social life and different societal contexts, such as work, education and politics.[22]

Lippert-Rasmussen defends this conceptual move on two grounds. He notes that social salience and paradigmatic cases of discrimination (e.g. sex discrimination) seem to go hand-in-hand. But more importantly, when a socially non-salient group (e.g. people with an uneven number of siblings) is discriminated against, the relevant harms attributed to being subjected to discrimination do not seem to necessarily follow. Treating people with an even number of siblings differently from those with an uneven number, he would argue, will not constitute objectionable discrimination due to the idiosyncratic nature of such treatment. Given this idiosyncrasy, being discriminated on the basis having an uneven number of siblings "will not seriously harm the disadvantaged party in the great majority of cases" (Lippert-Rasmussen 2014, 33). Because such traits are perhaps not significant for one's identity, it will not result in diminished self-respect or otherwise significantly affect their sense of self in the long run. Interestingly, Lippert-Rasmussen holds that it may constitute a case of unjust differential treatment, but not objectionable discrimination *per se*[23]. (Ibid., 33–34.)

By contrast, Eidelson maintains that we have no reason to limit the scope of wrongful discrimination to instances involving socially salient groups. Specifically, he notes that Lippert-Rasmussen's account seems to rest on an unintuitive externalist presupposition about the connection between the social significance of a trait and the discriminatory treatment:

> It would be very strange […] to say that whether [one] discriminates hinges on whether his idiosyncratic partiality happens to align with a distinction that is socially salient in his society at the time, although for reasons wholly unrelated to him or his partiality. That will bear on the consequences of his actions in many cases, of course, in ways that are potentially important to its moral status. But the idea that it determines *whether he discriminates* conflicts with what seems to me a basic "internalist" intuition about the concept. If we know everything about how X treats Y and Z and why he treats them as he does, we should know whether he is discriminating on any given basis in so doing. (Eidelson 2015, 28; italics added.)

[22] Lippert-Rasmussen concedes that it is difficult to say exactly when a group is socially salient. However, he does elaborate that social salience does not necessitate that a group sharing P has (i) many members (although this is often the case) or that (ii) there are many contexts in which P is of perceived importance. There may, for example, be only a few contexts in which P is of significance but is socially salient in that it has a substantial effect on how people interact in *those* contexts. (Lippert-Rasmussen 2014, 31.)

[23] Lippert-Rasmussen explicitly states that there "are cases involving disadvantageous differential treatment that we would probably consider discriminatory if the relevant groups were socially salient" (2014, 33).

To rephrase, it seems implausible that two otherwise identical acts would differ in their moral objectionability solely on grounds that one concerns socially salient groups while the other does not. This would be analogical to claiming that whether X's intentional communication of false information to Y counts as lying depends on which group Y belongs to. In this sense, the social salience account seems flawed in its unwarranted stipulation that discrimination on the basis of socially salient traits is the only form of possibly objectionable *discrimination*. (Eidelson 2015, 28–30.) Racist and sexist discrimination may, according to his account, nevertheless "command more of our attention because they are widespread" or because they bear a loaded cultural meaning, but this does not warrant the conclusion that similar treatment on the basis of even seemingly arbitrary traits could not fall under the category of (wrongful) discrimination overall (Ibid., 30).[24]

Refraining from limiting the category of wrongful discriminatory acts to those involving socially salient groups, I would argue that Eidelson's view fares better in comparison to Lippert-Rasmussen's. It recognizes instances of wrongful discrimination that are rare or uncommon, but nevertheless tick all the boxes with respect to what kinds of *acts* and *policies* are (or could be) discriminatory – and perhaps morally objectionable – were they to involve such groups. Still, his account does not undermine the moral significance of paradigmatic cases of discrimination (e.g. racist or sexist discrimination) as they may be conceived distinctively morally significant due to their prevalence in society, and thus more deserving of our attention. Having considered the role of social salience in the moral evaluation of discrimination, I will next consider three types of disrespect-based accounts of discrimination.

1.2.2. *The Mental State Theory*

Proponents of the MST hold that "discrimination is wrong when, and because, it is motivated by disrespectful mental states" (Beeghly 2017, 86). This view reflects the notion that there seems to be something wrong with treating people differently when this is explained by disrespectful beliefs about individuals of some group. Much of racist discrimination, for example, plausibly involves malicious beliefs about people of color, which then explain why they are discriminated against. The MST is considered, albeit not outright defended, by Larry Alexander who says the following:

> When a person is judged incorrectly to be of lesser moral worth and is treated accordingly, that treatment is morally wrong regardless of the gravity of its effects. It represents a failure to show the moral respect due the recipient, a failure which is by itself sufficient to be judged immoral. (1992, 159.)

[24] A similar argument is also made by Thomsen (2013).

While it is surely correct that blatant instances of discrimination may be motivated by beliefs about the inferiority or lesser worth of those discriminated against, "there is little reason to believe that disrespectful mental states motivate every single case of wrongful discrimination" (Beeghly 2017, 89). Refusing to hire women on the basis that they are more likely than men to take parental leave, would not be identified as instances of wrongful discrimination if motivated by business interests, for example. Furthermore, people may also discriminate due to unconscious, implicit cognitive biases. In fact, a wide range of literature on the subject suggests that many individuals exhibit implicit racial bias even if they explicitly endorse anti-racist beliefs (Kelly & Roedder 2008). Biased actions unmotivated by disrespectful mental states may be morally objectionable, however, given that they may further systematic disadvantage faced by vulnerable groups (Eidelson 2015, 108.) For example, hiring processes that systematically discriminate against women undermine their competence and equal right to work, even if those processes are not motivated by malicious intent. Accordingly, the MST seems to fail to satisfy Beeghly's identification and explanation conditions (cf. Beeghly 2017).

1.2.3. *The Expressive Theory*

Perhaps, then, the wrongness of discrimination stems from what the act is taken to *express*. Proponents of ET posit that "discrimination is wrong when, and because, it expresses a disrespectful social meaning" (Beeghly 2017, 87). Thomas Scanlon considers the idea that discriminatory decisions could be "objectionable because they involve a kind of insult— an expression of the view that certain people are inferior or socially unacceptable" (2008, 72). Deborah Hellman builds upon this notion, arguing that discriminatory actions and policies express a social meaning that demeans those whom are discriminated against (2008; 2017). For an act to demean a person or a group, Hellman says, it must satisfy two conditions: it must express "that a person or group is of lower status" and the discriminator "must have sufficient social power for this expression to have force" (2017, 102). The former condition is met when an act is demeaning in virtue of *respect-conventions* pertaining to some social context or culture. For example, segregation on the basis of ethnicity and gender bear "a loaded meaning" in many cultures, prominently due to historical oppression of those groups, as well as past and present inequality (Hellman 2008, 32). Racist and sexist discrimination bring those injustices to the fore and thereby denigrate those affected. The latter, what Hellman calls "the power dimension" is a necessary condition for the former to apply (Ibid.). That is, an act cannot demean someone in the absence of relevant social power held by the discriminator[25]. Discriminatory acts conducted by the

[25] In Hellman's words, "[i]solated actions by persons without power are insulting […] but they aren't demeaning" (2017, 104). Notably, whether an act does satisfy the first condition – whether it expresses that Y is lower status than Z – is dependent on neither X's intent to discriminate (cf. the MST), or the effects the act has on Y. What matters is only whether the discriminator, X, has "the requisite capacity to demean", understood as social power (Hellman 2017, 104). Indeed,

government and institutions with significant power over those they govern are thus at least *more* demeaning than those conducted by private actors, although acts conducted by the latter may be wrongful as well in many cases.

However, given that enacting policies of "indirect" discrimination will perhaps not always be regarded as conveying a demeaning message, Hellman's account possibly fails to satisfy the identification condition. Even facially neutral policies (e.g. height and weight requirements) may disproportionately exclude and burden different demographics and restrict their opportunities or access to goods and services. Thus, it may be possible to find instances where such requirements *would* constitute wrongful discrimination albeit no pejorative social meaning is conveyed in enacting them. (Beeghly 2017, 90.) Eidelson finds Hellman's account problematic in that acts are rendered disrespectful if and only if they are so "by the lights of the operative social norms" (2015, 88). Specifically, he emphasizes that intention seems to be morally significant for the moral evaluation of acts. One may fail to heed to respect-conventions by accident or due to ignorance, for example. Thus, while an act can be considered *conventionally* disrespectful, it may not nevertheless involve a disrespectful or malicious belief, or even an unconscious disregard for the equality of some individuals in relation to others. (Eidelson 2015, 87–88.) Furthermore, as acts will demean (in Hellman's sense) only those groups which have suffered from historical oppression or systematic disadvantage, the account also seems to limit the scope of wrongful discrimination to discrimination against socially salient groups[26]. Thus, Eidelson's has two objections against Hellman's account: On the one hand, the account may be overly strict due to the determinative role of respect-conventions and render innocuous acts disrespectful and thereby wrongful. On the other hand, it simultaneously risks overlooking idiosyncratic instances of genuinely disrespectful discrimination, such as treating people with an uneven number of siblings as if they were categorically lesser in value, just in case there is no loaded meaning expressed by such conduct.

1.2.4. *The Deliberative Failure Theory*

Finally, the "[d]eliberative theory of wrongful discrimination" holds that "discrimination is wrong when, and because, it manifests deliberative failure" (Beeghly 2017, 87). A version of DFT has been presented by Benjamin Eidelson, who holds that "acts of discrimination are intrinsically wrong when and because they manifest a failure to show the discriminatees the respect that is due to them as

Hellman (2017, 103) likens demeaning to ordering (in the sense of giving orders) in that both necessitate an asymmetry in power between people.

[26] What groups happen to be socially salient may, nevertheless, vary even in this view. However, Eidelson's point is that social salience does not matter for the moral evaluation of a discriminatory act altogether – socially salient groups may only be (actually) discriminated against more frequently and in a wider range of contexts.

persons" (2015, 73). Eidelson is talking specifically about so-called *recognition respect*[27] for persons, understood as "respect for someone's standing as a person" (Ibid., 74–75). Manifesting this sort of respect consists in affording appropriate weight to the equal value of persons and their autonomy when deliberating whether to act in one way or another with respect to that person. This notion draws from Immanuel Kant's categorical imperative, according to which one must treat persons as not only means but as ends in themselves (1785/1996). Recognizing the aforementioned elements of personhood "gives us reasons to deliberate *and* act in certain ways where [that person] is involved" (Eidelson 215, 78; italics added). A failure to do so – that is, a deliberative failure – manifests disrespect for a person's standing *as a person* whom adequate respect ought to be afforded. (Ibid., 75–77.) In this sense, disrespect does not ultimately consist in a positive attitude or belief (cf. MST) or a meaning expressed by an act in light of some respect-conventions (cf. ET) but it is, rather, a negative aspect of deliberation – a *lack* of respect.

The notion of recognition respect for personhood intertwines both the aspect of deliberation and that of action. The relevant object of moral evaluation is taken to include both the deliberative process (e.g. motives, intentions, inferential structure) and the act itself which follows this deliberation (2015, 77–78). Thus, a judgment as to whether an act constitutes morally wrong discrimination should concern not only the motives or the consequences of the act separately (i.e., as "thinly" described). Rather, the judgment should concern an act as it is *thickly* described; "by reference to how it is arrived at by the agent" (Eidelson 2015, 77). It is an agent's deliberation or motivation in conjunction with the act that follows it, in Eidelson's view, that we are concerned with in evaluating whether some act of discrimination is morally wrong. Differences in *how an agent comes to act* may constitute differences in the severity of the moral wrongness of that act.[28] Perhaps this is why one might hold that intentional manslaughter – i.e., murder – is *worse* than accidentally killing someone, even though both are arguably wrong. In Eidelson's view, if we fail to consider the act as it is thickly described, we overlook a significant aspect regarding the morality of that act.

Now, recall Eidelson's critique of Hellman's view; that it seems problematic to think that operative social norms should fix what disrespect consists in. Eidelson states this point explicitly

[27] Recognition respect is a concept coined by Stephen Darwall. Recognition respect, he says, "is a kind of respect which can have any of a number of different sorts of things as its object and which consists, most generally, in a disposition to weigh appropriately in one's deliberations some feature of the thing in question and to act accordingly" (Darwall 1977, 38). Recognition respect is distinguished from *appraisal respect*, which "consists in an attitude of positive appraisal" of some "person either as a person or as engaged in a particular pursuit" (Ibid.).

[28] Thick descriptions (e.g. 'selfish') can be understood as combining evaluative (e.g. 'wrong') and non-evaluative descriptions (e.g. 'conducting an act on the basis of self-interest'), whereas thin descriptions are more general evaluative terms (e.g. 'wrong'). Views supporting the irreducibility or separability of thick descriptions to their components are called inseparabilist views, whereas those claiming such separation is possible are called separabilist views. (Väyrynen 2019.)

by saying that "[t]o disrespect someone is to fail to take account of the normative significance of some facet of her moral standing; and it is just not up to a culture to decide what constitutes such a failure" (2015, 86). Going a different a way, then, he distinguishes between what he calls *basic* and *conventional (dis)respect*. (Ibid., 84–85.) Conventional disrespect is expressive, conceptually similar to Hellman's notion of demeaning, and is manifested in acts that violate respect-conventions persisting in some social setting (e.g. context, culture, and so on). Basic disrespect, however, does not depend on social conventions in the way that some social meaning would determine whether an act exemplifies basic disrespect. It is a lack of (recognition) respect; a failure to hold an appropriate attitude towards another person and to regulate one's actions accordingly. An integral difference between these concepts is that basic disrespect can only be attributed to thickly described acts, whereas conventional disrespect can be predicated of those thinly described. (Ibid., 84–90.) This also explains why both the MST and ET fail from Eidelson's perspective: they both conceive disrespect only in terms of thinly described acts and thereby overlook a morally significant aspect of those acts.

Differences withstanding, basic and conventional disrespect are, nevertheless, closely connected. As respect-conventions may determine what acts or beliefs can cause harm to individuals in a particular culture (e.g. shame as a psychological harm), they also partly determine what deliberative requirements a person must meet in acting within that culture. As Eidelson says, "it is not the convention itself that determines that my act is basically disrespectful, but rather my disregard for the harm that my act may cause *in light of* the convention" (2015, 85). Nevertheless, if the respect-conventions are not epistemically available for an individual – perhaps, he is unfamiliar with them – there will be no basic disrespect involved when he violates them, although his actions could be deemed conventionally disrespectful[29]. Showing conventional disrespect, then, should be understood as a contingent wrong, whereas basic disrespect is intrinsically so. For example, the thinly described act of spitting on someone is contingently wrong because one may do so by accident. The thickly described act of spitting on someone due to racist beliefs is wrong in every instance because it manifests a failure to recognize one's personhood. (Of course, the *fact* that spitting on someone involves racist beliefs is in itself a contingent state of affairs.) Notably, as basic and conventional disrespect are conceptually distinct, a given act can manifest either one of these when described thickly, or it might manifest both. (Eidelson 2015, 84–90; forthcoming, 20.)

Eidelson's account also has the wider implication that all respect-conventions may not in fact provide a conclusive case against the alleged wrongness of some acts, policies and practices.

[29] Assuming that he does not fail to recognize other persons as equal in value and as autonomous individuals, as is required by basic respect.

"Distinguishing convention-dependent [i.e., conventional] and convention-independent disrespect", he says, "allows us to see that our conventions about respect are *themselves* proper objects of moral assessment" (Eidelson forthcoming, 49). In other words, if there are situations in which some ethical principles, goals or attainable benefits justify a violation of some respect-convention, and where this can be done without clashing with the demand for recognition respect, we may rightfully ask whether that respect-convention grounds a fundamental moral requirement in the first place. Given that social conventions vary from context to context and from culture to culture, it seems correct that the moral demandingness of these conventions are negotiable, and that they themselves can be assessed from the point of view of ethics.

Notably, Eidelson also employs the concept of *contempt* to denote another type of disrespect. Contempt "involves a knowing refusal to take account of [a person's] moral standing" even "in the face of a minimal sort of recognition of it" (Eidelson 2015, 107). Thus, while basic disrespect consists in a failure to recognize a person's value, equality and individuality in relation to others, the concept of contempt captures cases in which an agent recognizes the moral demand and respect an object (e.g. a person or an animal even) imposes on him, and yet *refuses* to heed to this demand. Both basic disrespect and contempt will fall under the general category of disrespect, nevertheless, as a they are both failures to deliberate *and* act accordingly. Regarding the degree of moral wrongness between these two, Eidelson states that "it is especially bad to deny or refuse to recognize people's value as persons, as compared to simply *failing* to recognize it" (Ibid., 106).

The DFT, as presented by Eidelson, has the benefit of accommodating the general notions that are integral to both the MST and the ET. Discrimination on the basis of a belief about moral inferiority of some person will manifest disrespect, as such belief will involve a failure to recognize her equal value in relation to others. Similarly, if an act demeans some group – expresses conventional disrespect – we may often assume there is also basic disrespect involved, although there may be cases where this is not the case. But in any case, the moral evaluation of any discriminatory act should concern the deliberative process in conjunction with the act. Taking (dis)respect as a thick ethical concept, the DFT seems to both identify and explain instances of wrongful discrimination with more rigor than the MST and the ET (Beeghly 2017, 91).

1.3. Intrinsically Wrongful Discrimination

Eidelson's theory is a pluralist one, meaning that it allows for both intrinsically and contingently wrongful instances of discrimination. According to Eidelson's theory, what central cases of intrinsically wrong discrimination have in common is that, in addition to being discrimination in some

dimension W (in the non-moralized sense), they are cases in which individuals are discriminated against also "*in the dimension of respect for personhood*" (2015, 91). When discrimination is intrinsically wrong, it is because some person is not respected *as* a person because she differs in some manner from some other person. In more formal terms, X's reaction to P-wise differences among subjects figures into the explanation of why the discriminator fails to respect the disadvantaged individual or group. Racist discrimination, for example, is intrinsically wrong in that X's regard of Y race-wise leads X to disrespect Y's standing as a person, while X does not disrespect some other person, Z, because of how X regards Z race-wise. Wrongful discrimination, thus, involves *comparative* disrespect: a failure "to respect someone as a person when one does not (or would not)" fail to do so "in the case of an actual (or counterfactual comparator)" (Ibid.). Importantly, the notion of comparative disrespect is employed to distinguish genuine cases of intrinsically wrongful discrimination from cases in which *all* people are treated with disrespect but, nevertheless, not discriminated against. If an agent fails to recognize the personhood of all those subjected to his treatment, this may be immoral or unjust, but not discriminatory.[30] (Ibid., 74, 91–92.)

What is it, then, to which one fails to give appropriate weight when a deliberative failure occurs? Eidelson's theory is built upon two notions concerning what he takes to constitute the (moral) standing of a person: one's equal worth and her autonomy. Eidelson argues that with respect to these two, disrespect can manifest as a failure to recognize "a person's moral worth" and/or "her autonomy" (2015, 95). A deliberative failure of the first sort is a violation of what he calls the *interest thesis*:

> To respect a person's equal value relative to other persons one must value her interests equally with those of other person's, absent good reason for discounting them (Eidelson 2015, 97).

To satisfy the interest thesis is to respect the *equality* of a person in comparison to others. The interest thesis entails that respect for equality necessitates that an agent values a person and her interests, not only acknowledges them. In deliberating whether to perform some act with respect to a person, one must take into account that person's interests as reasons for and against that act. (Eidelson 2015, 97.) A person's "claim to presumptively equal consideration that is rooted in her equal value", Eidelson states, "is what sets the baseline of impartiality from which various differences may or may not warrant deviation" (Ibid., 98). That is, if there is no additional justification for not giving appropriate weight to a person's interests, a discriminatory act will be intrinsically wrong. Notably, that means

[30] As touched upon briefly above, it is contingent *whether* a discriminatory act in the non-moralized sense (i.e., thinly described) manifests basic disrespect. When Eidelson talks about intrinsic wrongfulness, he seems to imply that the kinds of thickly described acts that are wrongful because they manifest basic disrespect are wrongful in every instance and regardless of their consequences.

that irrespective of whether Y suffers from some harm due to this treatment or not, such discrimination will be wrong if one of the requirements stated in the interest thesis are not satisfied[31]. There are two ways this can happen: (i) X may fail to recognize and account for Y's interests in deliberation and subsequent action, or (ii) even if X does fail in this regard, he may fail "to make a serious effort to ascertain a person's interests accurately" (Ibid., 101). As the interest thesis requires that one actually values another person's interests, this poses a moral demand for one to try to *ascertain these interests*, given that this can be reasonably done[32].

A deliberative failure of the second sort – a failure to recognize a person's autonomy – will occur if a discriminator fails to satisfy one or both of the following conditions:

(Character Condition)	X gives reasonable weight to evidence of the ways Y has exercised her autonomy in giving shape to her life, where this evidence is reasonably available and relevant to the determination at hand; and
(Agency Condition)	if X's judgments concern Y's choices, these judgments are not made in a way that disparages Y's capacity to make those choices as an autonomous agent. (Eidelson 2015, 144.)

To satisfy these conditions is to respect the *individuality* of a person (by respecting her autonomy)[33]. According to Eidelson's view, alongside equal moral standing, autonomy is a "constitutive attribute of personhood" (2015, 162). The conditions reflect two distinct elements of autonomy: it is here understood both as deliberative agency, the capacity to make choices and reflect upon those choices, and as a condition that is actualized, being a "cumulative product of those choices" (Ibid., 141). Thus, what is integral to respecting a person's individuality will concern attending to the way a person defines herself and affording weight to the fact that this person is capable and free to make her own decisions. (Ibid., 145.)

How might violations of these conditions occur, then? Eidelson approaches the question by considering decisions that are made based on generalizations. Generalizations can take the form of what are called *generics*. They are "general claims about kinds and categories" (e.g. 'tigers are

[31] For example, passing a racial segregation policy will constitute intrinsically wrongful discrimination against people of color even if that policy is never enacted in practice.

[32] What is understood as "reasonable" in ascertaining a person's interests remains rather vague. In explicating his view about respect for autonomy and the role of information, he states that a decision-maker should not, at least, "discount information that (1) appears to reflect a person's autonomous choice and (2) is not less available or less probative than other information that one *does* take into account" (Eidelson 2015, 156).

[33] Eidelson understands individuality, roughly, as mental separateness and having one's own will. Individuals are capable of forming their own intentions and they "have their own wills of the kind that make for being a person in the first place" (forthcoming, 36).

striped') or claims about dispositions or probabilities concerning them (e.g. 'ducks lay eggs') (Leslie 2012, 355). Generics are neither universal statements nor statements that specify some individual because they lack a quantifier such as 'all' or 'some' (Beeghly 2018, 689). Eidelson takes that generics can be disrespectful when they manifest an attitude towards some group people "that is incompatible with fully recognizing them as beings of equal value" (2015, 129). For example, literally dehumanizing utterances (e.g. 'Jews are cockroaches') do not recognize the moral standing of members of some group as persons[34]. Generalizations may, however, manifest disrespect in other ways also, namely, by violating the Character and Agency Conditions.

On the one hand, one may disrespect a person's individuality by discriminating *solely* on the basis of information about some reference-class. By ignoring evidence that reflects an individual's own commitments, aspirations, the choices she has made in authoring her own life, an agent fails to treat her as an individual. (Eidelson 2015, 145–146.) For example, denying a former felon a job because of his criminal history may be disrespectful of his individuality. If other information, such as his determination to turn his life around, is excluded from consideration, his full character is not given sufficient normative weight in deliberation. This is a violation of the Character Condition: a group-generalization is taken to override ways in which a person has manifested her capacity for autonomous agency. Importantly, this is the case regardless of whether that generalization applies, because the moral violation concerns the deliberative process leading to an act (i.e., the thick description of an act)[35], and not merely the truthfulness of the propositional content of a belief. The generic statement 'former felons are more likely than others to commit crime' may be true, but as former felons do not form a homogenous group of people, individualizing factors should be considered alongside this information. Accordingly, Eidelson maintains that there is a difference between treating people on the basis of false beliefs and engaging in wrongful discrimination. (As I will argue, this distinction is also relevant for the moral evaluation of algorithmic discrimination.)

In Eidelson's account, whether and to what extent an individual *identifies* with some trait or group is also relevant for deliberative success, regardless of whether that trait is sensitive or legally protected. One's gender or race, for example, should not be understood as entirely irrelevant to decision-making. This goes against what some theorists have argued. The wrongness of racial

[34] For an in-depth examination of racial generics from the point of view of philosophy of language, see Langton et al. (2012). They find that characteristic generics involving racial characteristics should be rejected on the basis that they are both "false, and socially problematic" in that they portray a "social artefact as racial essence" (Ibid., 765). In addition, they hold that one should refrain from using statistical generics – even true ones – as they metalinguistically imply a characteristic generic, and thus produce epistemic and moral harms in the "invocation of racial natures" (Ibid.).

[35] It should be noted, however, that the Character Condition can be violated even in cases where X does not rely on generalizations. In fact, X could, in deliberating, afford weight *only* to the autonomous choices Y has made. However, if X fails to recognize some morally salient choices that are relevant to how Y's character and to how she should be treated, X may fail to respect Y as an individual. (Eidelson 2015, 155.)

discrimination, for example, is sometimes taken to be grounded in the notion that it is *irrational*. One's race bears no relevance to decision-making procedures and, as such, it should not enter into consideration. (Altman 2016, Sec. 4.1; see also Flew 1990.) Eidelson argues, by contrast, that "our unchosen traits may often be relevant to judgments about us as well", which is why the notion of respect "does not forbid affording these their epistemically appropriate weight" (2015, 152). This does not mean that one needs to explicitly *endorse* rules, practices and structures that depend on perceptions of race and racial categories in society, however (Eidelson forthcoming, 44). But to respect one's (exercise of) autonomy, attention how a person engages in self-presentational behavior – how she performs her gender or race – may even be required in some cases. (Ibid., 155–156.) Say an individual knowingly suppresses the perceptible salience of her gender by presenting oneself in ways that do not conform to stereotypical traits associated to that gender. In such a case, it would seem disrespectful to *not* consider that aspect of her behavior in relation to her gender when trying to appreciate her as an individual. Thus, for Eidelson, "appreciation of someone's autonomy [as character] seems to call for *including* more information, not *excluding* relevant information from consideration" (forthcoming, 40).

On the other hand, by relying on generalizations one may disrespect a person by failing to recognize her standing as an individual capable of making autonomous choices. Y's autonomy goes unrecognized if X views her as an object whose behavior and future performance is "determined by statistical tendencies" (Eidelson 2015, 148). This is a violation of the Agency Condition. To respect a person's autonomy, X's predictions about Y "must be predictions about how she will exercise her autonomy, rather than tacit denials that she *has* a full measure of such agency" (Ibid.). Again, the comparative aspect is significant here: X may well undermine everyone's autonomy in this way if all people are viewed *equally* as determined by traits attributed to some reference class (e.g. gender). However, X only discriminates against Y if Y's autonomy goes unrecognized and Z's does not. (Ibid.) Believing that women are driven by their emotions rather than by rational reflection, and that this does not apply to men, is to regard women as less capable of exercising their agency, for example[36].

To put it briefly, if the Character Condition focuses on how a person's choices have made her the individual she is, the Agency Condition is more future-oriented in that it requires recognition of her continuous capacity to author her own life (Eidelson 2015, 159). Notably, this view also accounts for *unconscious* generalization: The conditions that must be met for one to respect another as an individual are requirements to which "[c]onscious awareness does not seem to be of

[36] Eidelson explicitly refrains from considering the question of whether holding such a belief – regardless of acting on that belief – would be disrespectful in itself. Questions relating to this are considered in detail in the literature on epistemic discrimination (cf. Fricker 2003, Puddifoot 2017).

any basic significance" as they concern an agent's deliberative process, not conscious belief-states (Ibid., 165). One question persists, still: could partiality in decision-making be justified in some circumstances? Discriminating on the basis of personal preferences is commonly taken to be a form of discrimination that should perhaps be considered at least *prima facie* permissible. Choosing with whom one wishes to form social bonds or preferring certain types of people in comparison to others in matters of love and intimacy, are instances of so-called preference-based discrimination.[37]

Eidelson argues that there are two ways that preference-based discrimination will be objectionable. Firstly, preference-based discrimination is disrespectful if one's preference is based on a generalization that itself manifests basic disrespect. Say a woman prefers being examined by a female doctor. She may have such a preference not because she holds "a [generic] belief about what men are like" (Eidelson 2015, 121) and how they will conduct themselves but, rather, because she does not feel comfortable around male doctors, perhaps due to personal insecurity. This is not disrespectful discrimination, Eidelson argues, but it may be "regrettable" in the sense that it may (re)produce "systematic disadvantage to certain classes of people" if prevalent (Ibid., 125), such as male doctors' systematic exclusion from practicing certain fields of medicine. Secondly, preference-based discrimination is disrespectful if the preference involves a "tacit or explicit denial of the value of certain people" (Ibid.). One might, for example, hold that people of different ethnicities ought not intermingle. If this preference stems from the sole concern that intermingling would be against the interests of one's own ethnic group, such discrimination will manifest disrespect[38]. (Eidelson 2015, 116–117.) Likewise, "discrimination undertaken *in furtherance* of [an] ideal can itself show disrespect" (Ibid., 118). A person might endorse the view that people of different ethnicities ought not to intermingle for the sake of tradition, for example. While valuing tradition may not be disrespectful *per se*, furthering such an ideal may impose harm on those affected by such a norm. It may undermine their equal interest not to have that norm obtain and render their interests trivial. (Ibid., 118–120.)

Before moving onto contingently wrongful discrimination, one possible shortcoming of Eidelson's account should be considered. Lippert-Rasmussen (2019) points out that while Eidelson's notion of (dis)respect for equality includes a comparative element, this element is not fully fleshed out when it comes to (dis)respect for individuality. Lippert-Rasmussen entertains a hypothetical situation where a discriminator might in fact afford appropriate normative weight to evidence of how both Y and Z have exercised their autonomy and also respect the fact that they will continue to do so.

[37] A theory of the wrongness of discrimination is often hoped to accommodate this notion (cf. Lazenby & Butterfield 2017). Questions concerning discrimination and preferential treatment in the private and personal sphere are examined in Lazenby & Butterfield (2017) and Collins (2017).

[38] An example of this would be a white supremacist who prefers to associate only with caucasian individuals due to racist beliefs about people of color.

However, were the discriminator to respect Y's autonomy *to a lesser degree* than that of Z, it would seem inconsistent to not identify this as wrongful discrimination. Arguably, it involves a comparative difference in respect for the subjects' autonomy. In other words, it seems an agent might give *sufficient* and *reasonable* weight to how both Y and Z present themselves, but Y could be discriminated against in cases where (i) evidence concerning Z is afforded *more* weight or where (ii) Z's actions are taken to be *less* predictable in light of group-level information concerning her in comparison to Y (Lippert-Rasmussen 2019). Thus, it seems that Eidelson does not fully flesh out the notion of comparative disrespect in the context of autonomy.

Lippert-Rasmussen's critical argument seems to provide ground for slightly modifying the autonomy account. I do so by supplementing Eidelson's account with an additional condition. Let us call it the Equal Autonomy Condition[39]:

(*Equal Autonomy Condition*) X satisfies the Agency and Character Conditions equally with respect to Y and Z.

According to the Equal Autonomy Condition, an agent must not only respect the autonomy of those individuals subjected to his treatment to an appropriate degree, but she must do *equally so* with respect to each person, insofar as she is to avoid engaging in wrongful discrimination.

1.4. Contingently Wrongful Discrimination

Having examined what constitutes intrinsically wrong discrimination according to the modified deliberative failure account, I will next consider another side of Eidelson's pluralism, namely, his account of contingently wrongful discrimination. I will only provide a brief overview of his Broad Harms Argument, as these issues are considered in depth in chapter 4. In Eidelson's view, statistical discrimination is taken to be a type of discrimination that is contingently wrong, if at all. He defends this view by considering the case of racial profiling. (Eidelson 2015, ch. 6.) Racial profiling can be defined as "any police-initiated action that relies on the race, ethnicity, or national origin and not merely on the behavior of an individual" (Risse & Zeckhauser 2004, 136) It is important to note that it often involves using statistical evidence about criminal propensity among racial groups (e.g. arrest rates) in order to determine who should be subjected to (stricter) scrutiny. Risse & Zeckhauser's definition allows racial profiling to also denote practices where suspect descriptions contain information about individuals' race. In this study, I use the term (racial) profiling to denote practices of the former kind. Notably, it is not a necessary quality of racial profiling that it would rely *only* on

[39] The Equal Autonomy Condition is a slightly modified formulation of that presented by Lippert-Rasmussen in (2019).

information about individuals' race, which would be considered more straightforwardly objectionable as it would blatantly fail to respect subjects' individuality – it would constitute racist discrimination.

While arguments in defense of racial profiling emphasize it as an efficient method of locating and preventing criminal activity, objections to it stem from several distinct grounds. For example, it is argued that a when a racialized person is subjected to scrutiny as a result of profiling, she is not treated as an individual, but merely as a member of a reference-class. Similarly, it is taken to lead to unwarranted frisking and scrutiny of racialized individuals because race and crime are not causally connected. Some find fault in that people are discriminated against on the basis of traits they are not responsible for or have not chosen. Lastly, profiling is also held to express a message of inferiority and to exacerbate racism.[40] (Eidelson, ch. 6.2.; Hellman 2014; Lever 2007; Risse & Zeckhauser 2004.) These defenses and objections also apply to other forms of statistical discrimination and contexts (including algorithmic discrimination) and I consider them in more depth in chapters 3 and 4.

According to Eidelson, "profiling, and other practices like it, are very often *contingently* bad because of their *conventional* meanings" rather than intrinsically so, by manifesting basic disrespect (2015, 177). Reliance on non-universal generalizations in itself (particularly in the form of statistical information about groups) is not taken to be morally objectionable in his account; respecting a person's autonomy does not contradict with using predictive information about that person (Ibid., 145). There is, for Eidelson, a difference between treating people on the basis of evinced statistical generalizations as opposed to generics that are in themselves disrespectful (e.g. ones that deny the equal value of people). Thus, he refrains from adopting "a general skepticism about statistical generalization" but emphasizes the moral demand of attending to (reasonably available) information reflecting the relevant persons' autonomy and individuality (Eidelson 2015, 147).[41]

Nevertheless, Eidelson points out that thickly described acts involving statistical discrimination may yet be wrongful due to several reasons. He notes that

> one *could* adopt a racial profiling policy by virtue of a failure to afford equal weight to people's interests (e.g., in avoiding the burdens of law enforcement scrutiny), or because of statistical beliefs that are tainted by disrespect for certain people. Similarly, one could implement such a policy in a way that infringes the requirement to treat people as individuals: for instance, if race

[40] Arguments for and against racial profiling are often distinguished into two camps. On the one hand, consequentialist arguments generally focus on examining the costs and benefits of racial profiling. Non-consequentalist arguments, on the other hand, often approach the topic from a rights or fairness perspective. (Risse & Zeckhauser 2004, 132–133.)

[41] A harm-based objection against generalizations that Eidelson considers is that relying on and acting consistently on the basis of racial generalizations, for instance, may also further affect racialized people in a way that restricts their ability to exercise their autonomy in other social contexts. (Eidelson 2015, 154–155.)

> alone is taken into account, without regard for other aspects of a person's self-representational behavior. (Eidelson 2015, 174.)

In this passage, Eidelson suggests a crucial distinction between objections against statistical discrimination *in principle* and those concerning moral issues contingently associated with policies involving statistical discrimination. So-called *unalloyed* as opposed to *tainted* instances of statistical discrimination seem to call for different considerations[42]. The latter may be more straightforwardly objectionable as they might involve intrinsically wrongful second-order or direct discrimination. For example, statistical evidence could be unreliable or tampered with, tainted with disrespect and/or spurious[43]. The evidence might be enacted on in a selective manner, where subjects are scrutinized not on the basis of baseline propensities of their reference-class but, rather, due to police officers' prejudice or bias[44]. Yet as these disrespectful acts either obtain to different dimensions of conduct (e.g. arrest data collection) or do not exhaustively explain the possible wrongness of statistical discrimination, they should be considered separately from unalloyed discrimination, at least in theory.

Unalloyed discrimination, which is not accompanied by other disrespectful conduct and involves the use of *sound* statistical evidence, poses a more complex challenge. Indeed, critics of racial profiling often (at least implicitly) acknowledge the fact that "some statistically sound indicators exist as the residue of previous spurious ones" and the soundness of some of these indicators "is itself a product of previous and non-statistically justified discrimination" (Schauer 2017, 47). Eidelson's claim is that even if this is the case, when that statistical evidence is acted on in benevolence, this will not manifest basic disrespect. Elaborating on this, he says that

> unfairness in who is *selected* for the benefit or burden is probably not the heart of the matter. Nor is the problem simply that the exclusion employs a generalization or stereotype that will not prove correct in many cases. Rather, the moral case against the exclusion rests primarily on the harm it does, which requires close attention to the social meaning of the discriminatory policy, preexisting attitudes towards the groups affected, and the qualitative experiences of the

[42] The term *unalloyed statistical discrimination* is used in Lippert-Rasmussen (2006) to denote instances of statistical discrimination that are not tainted by other morally problematic contingencies. Theorists considering the subject of statistical discrimination often limit their examination to such instances (cf. Lippert-Rasmussen 2011, 55; Zarsky 2014, 1383). Lippert-Rasmussen, for example, states that "from the point of view of moral theory, it is of little interest to discuss statistical discrimination based on flawed generalizations" (2011, 55). As I understand this claim, it is not that it would be irrelevant to moral theory in general, but rather that one will have stronger, more straightforward objections to such conduct.

[43] One could object more straightforwardly to using unreliable method or evidence as a basis for discriminating. However, not all cases of statistical discrimination involve unreliable evidence. Thus, these two types of cases should be distinguished. (Lippert-Rasmussen 2007, 390.)

[44] Lippert-Rasmussen (2007, 390) notes that "[w]hile such selective use of statistical information is morally objectionable, obviously not all uses of information are selective in this way. In any case, the selective use objection never supports a conclusion of the form 'This kind of statistical discrimination is morally wrong.'"

> people who are discriminated against. These, I suggest, are the proper starting points from which to undertake a moral assessment of a practice of statistical discrimination that does not manifest disrespect. (Eidelson 2015, 221–222.)

He maintains that the moral evaluation of a practice involving statistical discrimination (e.g. racial profiling) should be approached from the perspective of whether the harms produced by it outweigh its benefits. In his account, statistical discrimination will be contingently wrongful if and when the so-called *broad harms* resulting from it overweigh its benefits. These harms include individuals' feelings of shame and denigration that follow from being scrutinized on the basis of sensitive traits, the stigma such practices reinforce, and the disrespectful conduct they may encourage.

1.5. Chapter Summary

In this chapter I have presented a view of what discrimination consists in the generic sense. It denotes differential treatment that results in comparative disadvantage for someone in some dimension (e.g. hiring). When direct, the treatment is explained by how an agent regards two patients to differ with respect to some property P (irrespective of whether they in fact have these traits). I offered a brief defense of the view according to which the set of properties on the basis of which discrimination can be rendered morally objectionable is not limited to socially salient ones, such as gender or race. I also examined other types of discrimination – namely, 'indirect', structural, and statistical discrimination. In Eidelson's view, contrary to many others, 'indirect' discrimination does not constitute a genuine type of discrimination. 'Indirectly' discriminatory acts are understood either as (un)conscious second-order (i.e., direct) discrimination or as instances of structural discrimination unmediated by an agent's regard of the discriminatees, which do not fall under genuine acts of discrimination. I also considered statistical discrimination, in which the presumed or evinced predictive connection between a predictor attribute P and a target attribute T serves as a ground for differential treatment.

After briefly comparing three conceptions of disrespect, I opted for the deliberative failure conception of disrespect which fares better in identifying and explaining instances of wrongful discrimination in comparison to the mental state conception and the expressive conception. Disrespect is here understood as a deliberative failure to account for the normative weight of a person's moral standing *as* a person. According to Eidelson's pluralist theory, instances of discrimination will be intrinsically wrong if and when an agent engaging in discrimination manifests disrespect towards the equality of persons – i.e., fails to recognize a person's equal interest in being treated in some way, and to actually value her interests by making an effort to ascertain them within reasonable limits. Discrimination will be intrinsically wrongful also if the discriminator fails to treat persons equally as

individuals, i.e., when the deliberative process involves an explicit or implicit failure to account for morally salient aspects of their individuality. Specifically, these are instances where use of group-generalizations (1) is coupled "with unreasonable *non*-reliance on other information deriving from a person's autonomous choices"; (2) contributes to a failure to recognize a person's autonomy *as capacity*; (3) is grounded in disregard for the discriminatees value as a person; or (4) does not reflect "an appropriately diligent assessment given the relevant stakes" (Eidelson 2015, 161).

Instances of wrongful discrimination that are not *intrinsically* so will wrong if and when the broad harms produced in engaging in a given practice will outweigh its benefits. Some instances of statistical discrimination, such as unalloyed racial profiling, fall under this category. They will be wrongful not by manifesting basic disrespect but, rather, because of the harm they produce, the meanings they express in virtue of respect-conventions, and the disrespectful conduct they sustain and encourage by reinforcing stigma associated to different demographic groups.

In the next chapter I will suggest that, as algorithmic discrimination can be understood as a form of statistical discrimination conducted on the basis of statistical profiles, questions related to the ethics of statistical discrimination will be relevant for this study as well. Accordingly, the discussion around racial profiling proves a useful starting point for considerations regarding algorithmic discrimination. Arguments against statistical discrimination in principle (and as I will show, algorithmic discrimination, by extension) will be considered and ultimately refuted in chapter 3. In chapter 4, I will suggest that Eidelson's Broad Harms Argument offers a more plausible account of why one would deem algorithmic discrimination wrongful, even in cases where they rely on statistically sound models and decision-makers have good intentions in using such tools. In the next chapter, however, I will examine ways in which AD might in fact manifest basic disrespect, when thickly described. I suggest that AD may be "tainted with disrespect" when it involves acts of disrespectful discrimination – specifically, discriminatory acts that take place in other dimensions of conduct, such as in making choices regarding data collection, as opposed to the dimension of generating algorithmic decisions in itself. Such tainted instances of AD will be analogous to cases of tainted racial profiling considered above. In instances of tainted algorithmic discrimination, the unfair outcomes of AD are not *exhaustively* explained by statistical facts, but also by second-order or direct discrimination, as well as other problematic conduct, such as carelessness and negligence, which may not constitute discrimination, but which may yet be morally objectionable.

2. Dissecting Algorithmic Discrimination (and Other Issues)

In this chapter, I identify distinct types of discrimination that may pertain in the development and use of ADSs by dissecting their design process and examining some key elements regarding their use. I examine four stages of the design, development and use of ADS: (1) selection of a target variable, class labels, and features (2) data collection and preparation, (3) modeling and model evaluation and (4) deployment.[45] In a nutshell, because AD ultimately relies on prediction and forecasting, automating decision-making processes requires determining what one is trying to predict and how. This is followed by collecting and pre-processing data that is deemed relevant to the determined task. The aim is "to construct a simple model with valuable use", most often "having high predictive accuracy" (Alpaydin 2016, 14). I focus on models constructed by using ML algorithms. ML enables that the system may be taught a "rule" (i.e., a decision function) from input to output by showing it examples of past decisions. ML may also be used to mine the data; to find novel, informative patterns within it, which provide actionable insight for decision-making. Once the learned model (or *classifier*) generalizes with sufficient accuracy to novel data, it may then be used to execute decisions, either autonomously or under the supervision or control of humans. Alongside, I examine how algorithmic bias may be introduced into the model. Drawing on existing literature on algorithmic bias and discrimination in AD, I locate two types of general risks for disparate impact. On the one hand, algorithmic decisions may discriminate against groups when the model captures existing inequalities reflected in data. On the other hand, design choices that require subjective deliberation and issues with human involvement in the use of ADSs may introduce bias into decision-making procedures, as well as mitigate or exacerbate existing bias exemplified by to those procedures.

I offer a preliminary account of different types of discrimination that may take place in the development and use of ADS, and which may be "layered" in a variety of ways. This layering and different senses of the term 'algorithmic bias', I suggest, partly constitute the perplexity of the phenomenon of algorithmic discrimination. Below I will also note that the views concerning distinct mechanisms of algorithmic discrimination distinguished by Barocas (2014; see Introduction) are in fact all correctly identified as discrimination, but they exemplify and comprise several distinct *types* of discrimination which may occur at distinct *dimensions* of conduct. None offer an exhaustive account of the ways in which AD may be discriminatory, however, nor of the ways in which they may

[45] The structure of my overview follows what is called the Cross Industry Standard Process for Data Mining (CRISP-DM), a common approach to designing and deploying of data science tools in the industry (cf. Kelleher & Tierney 2018, 56–67). The CRISP-DM is divided into six stages: (1) business understanding, (2) data understanding, (3) data preparation, (4) data modeling, (5) evaluation, and (6) deployment (Kelleher & Tierney 2018, 58). However, in my overview, these six stages are subsumed under four stages.

be wrong. I will suggest that wrongful algorithmic discrimination may be realized in multiple ways, partly due to the problem of many hands.

Some of the concerns expressed in the literature on algorithmic discrimination relate to discrimination in the development process or in the human supervised use of ADS. These include intentional biasing of the model at the stage of data collection or selective favoring of groups in using ADSs as informants, for example. While these instances of discrimination are problematic in terms of identification due to opacity of the design processes and due to the fact the relevant discriminatory outcomes often manifest only after deployment as disparate impact on some group, they may be more straightforwardly objectionable in that they "taint" the used statistical evidence in disrespect. I conclude by that not all discrimination in AD constitutes tainted statistical discrimination. So-called "unalloyed" algorithmic discrimination may yet reproduce existing inequalities by tracking structural inequity and disproportionalities between groups. That discriminatory acts are separated from other moral issues is integral, as they seem to require different moral considerations and countermeasures. I suggest that the moral evaluation of algorithmic discrimination ultimately involves questions about the moral permissibility of treating individuals differently on the basis complex algorithmically generated profiles, which may be understood as non-universal, statistical generalizations.

2.1. Target Variables, Class Labels and Features

A prerequisite for automating decision-making processes is to specify what information is of interest (to an organization) in a given decision-making context. This requires "identifying a business [or other] problem and then exploring if the appropriate data are available to develop a data-driven solution to the problem" (Kelleher & Tierney 2018, 60). In other words,

> data miners must translate some amorphous problem into a question that can be expressed in more formal terms that computers can parse. In particular, data miners must determine how to solve the problem at hand by translating it into a question about the value of some *target variable.* (Barocas & Selbst 2016, 678; italics added.)

Financial institutions, for example, use individual credit scores to stand for a specific trait – a person's "creditworthiness". For a data miner, a credit score is a prominent target variable in this context, values of which are to be predicted to support decision-making. By breaking down the target variable into a set of factors that contribute to it (i.e., the information that is considered in credit-scoring), the attribute of interest lends itself to formalization.

The underlying rationale of AD is that people with similar values with respect to the target variable should be treated similarly. This necessitates means to rank or differentiate between

those who are worthy of credit and those who are not, for example. Accordingly, a second design process involves specifying what values the target variable may receive, and the criteria according to which the value is determined. This process is called *class label* definition. Sometimes one may draw categorical distinctions between values (e.g. e-mails can be labeled as "spam" or "non-spam"). This is possible if the relevant categories are discrete, rather self-evident or uncontroversial, and mutually exclusive (e.g. an e-mail either is spam or not[46]). In cases where the variable can receive continuous values, however, one may need to apply a threshold to differentiate between outcomes (e.g. degrees of creditworthiness). Thus, while categorical distinctions allow for sorting novel data instances into discrete classes, applying thresholds to continuous values enables more fine-grained ranking of data. Defining the relevant class labels is integral for training a model via ML because it determines the categories under which the examples in the training data (i.e., previous human-made decisions) will fall. (Barocas & Selbst 2016, 677–680.)

Lastly, a data miner also needs to identify what information is needed in order to *calculate* the value of a target variable. Credit scores, for example, may be calculated using a number of factors, e.g., customers' income and existing debt. In a process called *feature selection*, the decision-makers and data miners preliminary determine "what attributes they observe and subsequently fold into their analyses" (Barocas & Selbst 2016, 688). This provides focus for subsequent data collection; e.g., if income is a predictor for one's credit score, income data needs to be collected from customers. The process may be iterated in the modeling phase as data-analysis may prove some features to be less significant for calculating the value of a target variable. So-called dimensionality reduction algorithms can be used to detect features in the data that are pivotal for predictions. When unimportant features are removed from the model, this will decrease the amount of input data needed and it may also improve the performance of the ADS. (Alpaydin 2016, 73–76.)

2.1.1. *Target Variable Bias*

Referring to the delicate process of determining a target variable, Barocas & Selbst argue that in the "necessarily subjective process of translation, data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes" (2016, 678). While there are limitations to how target variables can be defined altogether (e.g. it must be relevant to the decision to serve a practical purpose), some definitions may prove significantly worse than others, even in constrained contexts. Barocas & Selbst note that "hiring decisions made on the basis of predicted

[46] Of course, it is a different question whether and how well an e-mail can be *identified* as one or the other. The question of identification is, indeed, what ML is used for – data mining is used to identify factors that correlate with e-mail being spam, and then a classifier is trained to calculate whether an e-mail is spam or not.

tenure", for instance, "are much more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity" (2016, 680). The latter may be less susceptible to having a disparate impact due to its ability to track individual desert while the former may reflect discriminatory practices of those who have executed hiring decisions in the past (e.g. sex discrimination in hiring).

Target variables may also "leak", as Virginia Eubanks shows in her case-study on the Allegheny Family Screening Tool (AFST) used by Allegheny County officials to predict risk for child maltreatment in families. (2018, ch. 4.) She states that while "Allegheny County is concerned with child abuse, especially potential fatalities", only few such cases occur per year, which means a "meaningful model cannot be constructed with such sparse data" (Eubanks 2018, 143). Instead, Allegheny County opted to use two proxies as target variables: community re-referral and child placement. As a result, AFST did not actually predict child maltreatment *per se* – it predicted (i) which families will be regarded by the community as high-risk of child maltreatment, and subsequently reported to the officials, and (ii) which families will be regarded similarly as such by the agency, and subsequently have their children taken from them. (Ibid., 144.) What Eubanks finds out is that the model's predictions were effectively skewed by the fact that some families were continuously harassed by other members of the community; some families were re-referred to the system due to nuisance calls made to hotlines where one can report suspicions about child maltreatment. (Ibid., ch. 4.) In other words, the selected target variable(s) left room for the model to become skewed by racial and socio-economical prejudice, harassment and discrimination pertinent in the community.[47]

2.1.2. *Controversial Class Labels and Coarse Features*

As Barocas & Selbst note, "the definition of a target variable and its associated class labels will determine what data mining happens to find" (2016, 680). Data miners' interpretations of the target attribute – which are reflected in class label definitions – may introduce bias into the model in that the specified thresholds and criteria for class-inclusion may potentially disfavor some demographic groups. What makes an individual creditworthy, for example, is arguably open to interpretation, which makes the measurement of such a social variable difficult (e.g. how much debt makes one unworthy of credit?). Creditworthiness is also tightly connected to several fluctuating social

[47] Chouldechova et al. find the same problem in their case-study on the AFST and state the following: "One might be concerned that this outcome variable is simply a proxy for an unobserved or difficult to observe outcome of greater interest such as, say, severe maltreatment or neglect. Race-related differences in *reporting rates*, *screen-in rates*, and *investigations* may mean that the target variables are in closer alignment with the outcome of interest for some groups than for others. This is arguably an even bigger issue in criminal recidivism prediction where one is interested in re-offense but instead predicts re-arrest than in the child welfare context." (Chouldechova et al. 2018, 10; italics added.)

contingencies (e.g. income, job status, employment rate in a given area) that align with sensitive group-membership, which increases the risk of disparate impact on those groups. (Ibid., 679–680.)

A similar risk lies also in the process of feature selection. Here, "the coarseness and comprehensiveness" of the selected features may lead to disproportional rates with respect to how "different groups happen to be subject to erroneous determinations" (Barocas & Selbst 2016, 688). If the selected features are overly general, they could lead to drawing "distinctions between subpopulations" that will "fail to capture significant variation within each subpopulation that would result in a different assessment for certain members of these groups"[48] (Ibid., 689). This may result in higher rates of either erroneous or comparatively worse predictions for some groups. (Ibid., 688–690; see also Zarsky 2014, 1391) Thus, when data miners pick out which predictors to use, this may lead to adverse effects on certain groups through careless conduct but also via transferring existing inequality from other domains. When features that encode disproportionalities between protected groups are used in other contexts, this bias may lead to disparate impact on the same groups in these contexts as well. For example, when credit scores are used in making decisions regarding hiring, the fact that minorities and women are overrepresented in the low-end of these scores[49] may affect negatively on their assessments in employment contexts (Silva & Kenney 2017, 20).

2.2. Data Collection and Preparation

The way the target variable and class labels are defined, and choices as to what features are to be observed, guide the data miner in determining what type of data is needed for training a model. The next step, then, is to gather and prepare this data. Organizations may utilize the data they have generated as a side-product of day-to-day operations, but they may also rely on external data sources. (Kelleher & Tierney 2018, 93–94.) For example, police departments use data gathered from social networks and commercial sources, in addition to data from internal reports (Selbst 2017, 113). There are also many public and free-to-use data sets available online for different purposes, including research and commercial purposes[50]. Data also requires tedious preparation before it can be mined by using ML algorithms. It needs to be extracted from different sources, checked for quality, cleaned of inaccuracies and corrupt data, converted into a standardized format, and pre-processed before it can

[48] One should note that this is a distinct risk of disparate impact, while the mechanism is similar to the one identified in selecting the target variable. Features are variables used in the model, while the value of the target variable is what one is trying to predict. Credit scores may be used in some context as features, for example, while, in some other context, they are the object of prediction.

[49] Citron & Pasquale state that "critics have questioned the fairness of credit-scoring systems" due to their "disparate impact on women and minorities" in the U.S. (2014, 10).

[50] For a list of publicly available datasets see Stanford (2018).

be adequately utilized. Data may also need to be enriched by providing annotations, labeling it and reorganizing it under a novel categorical structure (e.g. spreadsheets).[51] (Kelleher & Tierney 2018, 60–61.) The prepared set of data is commonly separated into a training data set and a validation data set. The former serves a set of example decisions from which a ML program "learns" the model, and the latter set is used to test the trained model for accuracy and performance. (Alpaydin 2016, 155.)

2.2.1. *Data Collection Bias*

Several entry-points and sources of so-called *training data bias* are located at the stages of data collection and preparation. These comprise "[b]iases introduced due to the selection of data sources, or by the way in which data from these sources are acquired and prepared" (Olteanu et al. 2019, 13; see also Danks & London 2017, 4692–93).

A sub-class of training data bias, *data collection bias*, comprises problems related to the representativeness of the data in relation to the population one is interested in, and which may compromise the generalizability (or external validity) and accuracy of the model. A model trained on an unrepresentative set of training data (i.e., a biased sample) will not generalize adequately to the population, which can lead to differences in accuracy with respect to different groups. Conversely, the training data may in fact be representative of the population but reflect a current state-of-affairs that one might wish to change by deploying the model. That is, while there might be no significant distortion between the sample and the population, one might want "to ensure that [the model] does not maintain a morally problematic status quo" (Danks & London 2017, 4693).

Representativeness is a significant worry as marginalized groups may be over- or underrepresented in data sets used to train algorithms. For instance, 'Labeled Faces in the Wild', a data set used for training image recognition systems, contains 15000 images of faces with only 7 % of them belonging to people of color (Han & Jain 2004). This may ultimately lead to higher rates of misclassifications and mispredictions for these groups in AD. Given that readily available data sets are typically organized and pre-labeled, inherent bias in these datasets may also risk spreading when such datasets are used without due caution.[52]

Underrepresentation in data may be a consequence of lack of access as well. Predictive models used by public agencies may rely on datasets in which people in poverty are

[51] Data collection and preparation is estimated to consume around 80 % of the time spent on a data science project (Kelleher & Tierney 2018, 67).

[52] The worry of systematic underrepresentation of different demographics – minorities in particular – both in (training) data and at the institutional level of AI research and development (cf. West et al. 2019) has led scholars and developers to advocate diversity in AI as a way of ensuring inclusion-in-design. Jobin et al. find that diversity is considered a core principle of ethical AI in a majority of guidelines for developing ethical AI. Specifically, in their review, the principle of diversity is listed under the more general principle of justice and fairness, which is found in 68 of the examined 84 guideline documents. (Jobin et al. 2019.)

disproportionately represented, and this may be partly due to the limitations on the data the agencies have access to. The AFST, for example, utilizes data that is collected mainly from public agencies (e.g. child welfare and mental health services). As such, low-income families are oversampled in the data, and wealthier families who have access to private services are subsequently neglected in the data pool.[53] (Eubanks 2018, ch.4.) A related problem concerns so-called "dark zones" in data. Due to differences in how sub-populations engage in online services (e.g. frequency and fluency of engagement) or have access to them, some parts of the population will be underrepresented in available data. If limited access aligns with lower income, dark zones may lead to systematic exclusion of individuals and groups from data sets due to conditions of poverty. (Barocas & Selbst 2016, 684–685.)

2.2.2. *Data Preparation Bias*

The process of data preparation involves a risk of *data preparation bias*[54], which comprises distortions "introduced by data processing operations such as cleaning, enrichment [e.g., data labeling], and aggregation" (Olteanu et al. 2019, 14). Data preparation bias may affect both the internal and external validity of the model. (Ibid., 10–16.) Data cleaning (e.g. removing or correcting data items and entries) serves to improve validity but it may also "embed [one's pre-existing] beliefs about the phenomenon and the broader system into the dataset" (Ibid., 15). Data miners' decisions in data preparation may be influenced by *confirmation bias*, for example; a cognitive tendency to look for and rely on evidence that supports one's preliminary conceptions and existing beliefs about a given phenomenon (Nickerson 1998; Kaminski 2019, 1538). Essentially, this leads to similar problems as in feature selection: if certain types of data are removed from the training data set (e.g., radical values in the distribution of data points) some desirable features which would incorporate meaningful variance into the model may be excluded as well.[55]

Enrichment (e.g. labeling training examples) may "both exacerbate existing biases, as well as introduce new biases and errors" (Olteanu et al. 2019, 15; see also Barocas & Selbst 2016, 682–684). Data miners can label data manually themselves or rely on external pre-labeled datasets. In both cases, inadequate data labeling may risk discrimination by biasing "the resulting findings such that any decisions taken on the basis of those findings will characterize all future cases along the same lines" (Barocas & Selbst 2016, 681). The risk concerns supervised learning methods (see chapter 2.3)

[53] Notably, this does not mean that the model would not be accurate. The concern is, rather, that this kind of disproportionate focus in data collection may reinforce a negative cycle in which those individuals from whom the data is collected are also targeted by the system utilizing that data.
[54] Oltaeanu et al. (2019) use the term "data processing bias" instead of data preparation bias.
[55] Radical values and outliers in data are often removed explicitly to increase the model's performance (cf. Kelleher & Tierney 2018, 119).

in particular as the machine will infer a decision function from input data to a given output prediction based on training examples that have been manually labeled by data miners.

Three distinct concerns relate to data preparation. Firstly, data labels may be erroneous. In his study, Selbst finds that incorrect labeling of past crime data may result in disproportionate targeting in the use of predictive policing algorithms (i.e., over-policing):

> Results after arrest are often not updated. Thus, most research in crime statistics and most actuarially-driven criminal justice systems use arrest data as the best available proxy, even though arrests are racially biased and in other ways a poor proxy for crime. Even if post-arrest statistics were collected, a great number of cases end in plea agreements that do not reflect the crime the arrestee committed or was originally arrested for—thus, using these statistics would not solve the problem either. As a result, a majority of crime labels may be incorrect, whether describing a type of crime or the existence of one, and thus models will learn that people of color commit a higher percentage of "crimes" than they do in reality. (Selbst 2017, 133–134.)

In other words, incorrectly labeled training data skews the model, and subsequent actions taken on the basis of its predictions will be based on more or less spurious correlations.

Secondly, borderline cases in training data pose a risk for subjective biasing. It may be difficult to determine "which of the available labels best applies to a particular example", even when the class labels themselves might be rather "uncontested or uncontroversial" (Barocas & Selbst 2016, 681). On the one hand, some training examples may satisfy some criteria specified by a class label but not all, which introduces confusion as to which class an example should fall under[56]. On the other hand, insufficiently detailed criteria for class-inclusion might result in a situation where an example could be taken to fall under two distinct classes. (Ibid., 681–684.)

Lastly, data preparation bias may be an artifact of the ways the target variable and class labels are defined. Here, example labeling is not biased in the sense of being erroneous or a result of subjective skewing. Rather, the labeling process only inherits existing bias in the distribution of examples. Barocas & Selbst note that a "computer may learn to discriminate against certain female or black applicants if trained on prior hiring decisions in which an employer has consistently rejected jobseekers with degrees from women's or historically black colleges" (2016, 682). Thus, the data labels ("hired" or "not-hired") may be correct, but nevertheless reflect past discriminatory practices.

[56] For example, it is not self-evident how many missed credit card payments should render an individual unworthy of credit (Barocas & Selbst 2016, 681). This problem becomes even more pertinent when the categories in the data increase – e.g., when one has to look at other attributes in addition to missed payments, such as income, job status, and so on.

2.3. Data Mining, Modeling and Model Evaluation

The training data is used to "teach" ML programs; to detect and learn regularities and patterns in the data. Differing from more traditional computer programs where such rules and structures are hand-coded[57], ML programs can be understood as "general templates with modifiable parameters" (Alpaydin 2016, 24). Changing the value of a given parameter results in a change in the program's performance. By using a ML algorithm, the template's parameters can be automatically adjusted to fit the data in an optimal manner (i.e., to approximate a decision function). Adequately adjusted parameters contribute to efficiency in performing tasks, such as recognizing similar images. The "learning" in ML refers to the automated optimization of the parameters by using a learning algorithm which trains the program on a set of training data. (Ibid., 24–25.) ML methods, then, are different methods for "teaching" the program. They vary in the types of tasks they are suited for, what sort of data they require, and the amount of human involvement they necessitate. Choosing a given ML method and a learning algorithm requires balancing between the complexity of the model, its ability to generalize to new data, and its computational and practical tractability. (Danks 2014, 155–156.)

A first method called *supervised learning* is used when both the value of the input and the desired output are known, and the aim is to approximate a decision function from the input to the (target) output. Typically, supervised learning is used for classification and regression tasks, which involve automated "recognition or identification" of individuals or objects and building models "that can be used to predict the category membership of new individuals" (Danks 2014, 154).[58] A classification algorithm (i.e., classifier) could be taught, for example, to classify images of dogs and cats into appropriate classes. The algorithm searches for a "function that maps from the values of the input attributes to the values of the target attribute" within the data and produces a model "that implements this function" (Kelleher & Tierney 2018, 99). During training, the machine gradually reduces the error between the output value and the desired value to approach the optimal function (LeCun et al. 2015, 436). Importantly, supervised learning requires *labeled data*. For the algorithm to

[57] ML programs differ from earlier *expert systems* that rely on a hand-coded set of conditional rules and heuristics to transform input data to an output (Boden 2014, 91–92).

[58] Regression tasks can be approached either via linear or logistic regression. Linear regression is a method for finding the coefficients for some variable(s) and a dependent target variable. It is best suited for data with numeric attributes. Logistic regression is used when the dependent target variable has a discrete binary value (e.g. pass or fail), which means it is also suitable for classification tasks. (Kelleher & Tierney 2018, 114–120.) In general, classification and regression tasks differ only in that "classification involves estimating the value of a categorical attribute and regression involves estimating the value of a continuous attribute" (Ibid., 179).

find the most optimal function, the target attribute value (e.g. hired/not hired) corresponding to each input value needs to be included in the training data.[59] (Kelleher & Tierney 2018, 99–100.)

In *unsupervised learning* there are no defined target variables. Rather, the aim is to *mine* the data; that is, "to provide a general characterization of the full dataset" and to gain novel information about patterns within it (Danks 2014, 154). Consequently, there is no need for data labels. Unsupervised learning algorithms are used for estimating an underlying probability distribution in the data (i.e., density estimation) and to reduce complexity and the number of uninformative features in the model (i.e., dimensionality reduction; see chapter 2.1). They are also used to separate data objects into groups based on some measure of similarity (i.e., clustering) or for finding data objects or attributes that co-occur in the data (i.e., association-rule mining). Clustering algorithms[60] are used to provide information about what data objects are similar to another, while association-rule mining can help in predicting the occurrence of one data object based on the occurrence of another.[61] (Alpaydin 2016, 73–74; Danks 2014, 154–155; Kelleher & Tierney 2018, 164–165.)

Lastly, *deep learning* comprises an approach, which is not specifically a distinct form of ML. Rather, it is distinguished in that it involves using so-called deep artificial neural networks[62] (deep ANNs) which comprise multiple layers of non-linear processing and up to hundreds of millions of adjustable parameters. They are highly recursive and can learn more complex functions, and

[59] ML algorithms used in supervised learning "differ in the models they use, or in the performance criteria they optimize, or in the way the parameters are adjusted during this optimization" (Alpaydin 2016, 39). For example, the *k-nearest neighbor* algorithm can be used to find similar objects in the data. By presenting the algorithm some specific example, it will search for instances that are similar to the example in their attributes. (Ibid., 60.) *Support vector machines* perform classification and regression tasks by constructing a hyperplane between nearest data-points between classes (cf. Kelleher & Tierney 2018, 163). The larger the distance between these points, the smaller the generalization error.

[60] For example, the *k-means* algorithm is a widely used clustering algorithm (cf. Kelleher & Tierney 2018, 156–160). It separates a set of scattered data points into some predefined amount (*k)* of clusters. It calculates the mean values of those clusters, which in turn function as new means for each cluster, respectively. The process is iterated until clusters converge.

[61] Two methods are commonly named alongside supervised and unsupervised learning. *Semi-supervised learning* is a method mixing both supervised and unsupervised learning: Only a small amount of labeled data is presented to the machine in addition to unlabeled data. Semi-supervised learning is a method suited for domains where unlabeled data is more readily available in comparison to labeled data. (Kingma et al. 2014, 1.) *Reinforcement learning* is a method which involves no labeled data but, rather, feedback from a human observer. The machine explores different sequences of actions in a choice-environment and is provided with feedback (i.e., reward/no reward). It then to adjusts its future actions accordingly in an effort to maximize the reward. (Alpaydin 2016, 125–128.)

[62] The concept of ANNs stems from McCulloch and Pitts' (1943) seminal work in which they showed how networks of processing units – sometimes likened to biological neural networks – "could perform any Boolean operation (and, or, not) and, thus, any possible computation" (Franklin 2014, 16). ANNs are "composed of artificial neurons and synaptic connections, where each neuron has an activation value, and a connection from neuron *A* to neuron *B* has a weight that defines the effect of *A* on *B*" (Alpaydin 2016, 86–87). When *A* is active, the synapse will either excite or suppress *B* depending on whether the synapse is excitatory or inhibitory. The neurons are organized into layers with each neuron being connected to one or more neurons in the subsequent layer (in recurrent networks connections may obtain between neurons in the same layer or preceding layers, or reflexively when a neuron is connected to itself). Neurons in the input layer propagate the activation to so-called hidden layers that execute further processing. Finally, the activity in the network propagates to an output layer where the output value is calculated. The approximation of a target function (i.e., learning) consists in the reinforcement of weighted connections between neurons when activated simultaneously. (Alpaydin 2016, 86–90; Arkoudas & Bringsjord 2014, 52; Kelleher & Tierney 2018, 132–133.)

thereby discriminate more efficiently between relevant and irrelevant information. Because supervised deep learning requires large amounts of labeled data, deep learning is typically conducted unsupervised. (LeCun et al. 2015, 436.) Requiring large sets of training data but minimal human contribution, unsupervised deep learning is efficient in extracting novel patterns and regularities from data, making it an attractive method for data mining purposes. (Alpaydin 2016, 106–108.) However, it involves significant issues with model explainability and interpretability (see chapter 2.4.2. below).

2.3.1. *Algorithmic Processing Bias*

The modeling process involves a risk of *algorithmic processing bias*, which is a result of choices as to what algorithms are used in analyzing and processing the data. This stage of model design includes an iteration of the feature selection process where irrelevant features are excluded from the model either manually or by using dimensionality reduction algorithms. Thus, it also involves similar risks, e.g., excluding relevant information from the model and thereby reducing meaningful variance (see chapter 2.1.2.). Processing bias can be also introduced explicitly, as it may improve the systems performance. Biased statistical estimators are used to achieve more robust and reliable performance and problems with anomalies and noise in the data (e.g. extreme and erroneous values) may be tackled by using biased processing. (Danks & London 2017, 4693; see also Springer et al. 2018, 451.)

Processing bias may also be introduced in order to regulate an ADS's actions in light of ethical standards, even if this will limit its performance with respect to the main objective. Danks & London say that "an autonomous weapons system", for example, "might be provided with an ethical regulator that will not allow it to fire at perceived enemy combatants if they are near a UNESCO protected historical site" (2017, 4693). The ADS's performance is in this example limited by an overriding ethical (processing) principle, rendering the system's processing biased in a computational sense. This example serves merely to show that bias should not be understood as an indication of inherent error or (moral) flaw in the system's performance – it is also a way of correcting other biases (e.g. training data bias or issues with scarce data) or regulating an ADS's actions.

2.3.2. *Model Evaluation: Fit and Fairness*

Once the program has learned a model, it is typically tested on validation data for accuracy and performance. For example, if the ADS is supposed to execute a classification task, it will be evaluated on the basis of how well it recognizes novel data objects and sorts them into appropriate classes. There are two concerns relating to a model's accuracy and its generalizability to novel data. On the one hand, an overly accurate model will risk *overfitting*. Here, the model corresponds exactly or closely to the training data and may not generalize to novel data. In other words, the system has merely "memorized" the training data and will fail to perform in the face of novel instances. On the

other hand, an *underfitted* model will not perform well even on the training data. Such an algorithm has learned rules that are overly loose, which will result in inaccurate predictions (i.e., false negatives and false positives). At this stage of the design cycle, it is common that the model is tested and refined until it surpasses some threshold in accuracy. (Kelleher & Tierney 2018, 145–148.)

An accurate model is not necessarily a *fair* model, however. Given that the training and validation data are separate subsets of the overall data that has been collected, both will often contain the same biases. Thus, an algorithm trained on biased data may perform *accurately* on the validation data as well, which is why the model evaluation phase may also involve an evaluation of model fairness. Fairness evaluation involves applying some metric for fairness in decision-making and checking whether the ADS satisfies that metric. If it does not, designers can try to mitigate the amount of bias in the model's performance by revising choices that were made in the design process, such as data collection and preparation processes or choices regarding used algorithms, for example.

A wide range of definitions for algorithmic fairness have been offered in the literature (cf. Verma & Rubin 2018) and they can be distinguished roughly into three categories. Firstly, (i) statistical definitions for fairness define fair ML (or AD) in terms of statistical metrics, such as true positive and true negative rates, and false positive and false negative rates. A model may be evaluated on the basis of whether members of protected and unprotected groups (e.g. women and men) have an equal probability to receive a positive prediction (i.e., decision), for example. If they do, the model satisfies a common definition called *statistical parity*.[63] The second category of definitions, (ii) similarity-based measures, consists in ways of checking whether two otherwise identical individuals who differ only with respect to their gender or race, for example, will receive the same prediction.[64] One such definition, *fairness through unawareness*, requires "blindness" to sensitive attributes; the idea is that the AD process ought to not explicitly take into account the sensitive attributes of an individual. (Verma & Rubin 2018, 5–6.) Lastly, (iii) causal approaches to fairness comprise those

[63] Other definitions include *predictive parity*, *predictive equality*, *equal opportunity*, *conditional use accuracy equality*, *overall accuracy equality* and *treatment equality*. Predictive parity is satisfied if protected and unprotected groups have an equal positive predictive value (i.e., the probability that individuals who receive a positive prediction are members of the positive class). Predictive equality is satisfied if they have an equal false positive rate. Equal opportunity requires balance between groups in false negative rates. Conditional use accuracy equality requires that predictions are equally accurate with respect to members of both protected and unprotected groups in both the negative and positive prediction classes. Overall accuracy equality requires that overall accuracy of the ADS is equal between protected and unprotected groups. Lastly, treatment equality is satisfied when the misprediction rates (false positive and false negative rates) are equal between these groups. (Verma & Rubin 2018, 3–5.)

[64] Similarity-based measures also include definitions such as *causal discrimination* and *fairness through awareness*. A model does not exhibit causal discrimination if individuals who differ only in their protected group-membership yet have otherwise identical attributes receive the same predictions. (Verma & Rubin 2018, 5–6.) Fairness through awareness consists in the notion that similar individuals ought to receive similar predictions. "The similarity of individuals is defined via a distance metric; for fairness to hold, the distance between the distributions of outputs for individuals should be at most the distance between the individuals" (Verma & Rubin 2018, 6).

that "are not wholly data-driven but require additional knowledge of the structure of the world, in the form of a causal model" (Loftus et al. 2018, 4). They incorporate assumptions about the causal structure between sensitive attributes, proxies for sensitive attributes, and outcomes in the decision-making process. The idea is to identify and estimate the effect sensitive attributes have (directly or via proxy) on the outcome in the model, and to mitigate this effect if it is deemed illegitimate.[65] (Verma & Rubin 2018, 6–7; Loftus et al. 2018, 4.) For example, applicants' race may have affected their credit scores in the past due to discrimination. Given that past credit decisions may be used to predict new ones, the undesirable effect of race on past credit scores – e.g., the connection between the factors – could perhaps be mitigated or removed from the model to make it fair.

I consider fairness in more depth in chapter 3, where its relation to discrimination is considered. At this point it suffices to note that fairness evaluation by way of employing different fairness metrics is a way of identifying and mitigating possible bias in algorithmic systems.

2.4. Algorithmic Decision-making and User Bias

Once sufficient accuracy is achieved the model can be put to use. Importantly, use of ADS "in the wild" can vary with respect to the degree of human involvement and oversight. Two approaches are to be distinguished here: In a "human-out-of-the-loop" approach, the system scores individuals and enacts decisions autonomously without human interference or oversight. Conversely, in a "human-in-the-loop" approach the system acts as an informant, providing an output score based on the model. The decision and execution thereof, however, is left to a human decision-maker.[66] (Citron & Pasquale 2014, 6–7; Ethics Guidelines for Trustworthy AI 2019, 16.) Insofar as these approaches represent different ways in which AI and humans may interact in decision-making, they should also be accounted for in how discrimination may occur in real-life AD contexts. Indeed, several sources of so-called *user bias* can be located at the level of use and human-computer interaction. Two subclasses

[65] Causal reasoning approaches also include *fair inference*, *no unresolved discrimination* and *no proxy discrimination*. Fair inference involves defining some causal pathways from protected attribute to a decision as legitimate or illegitimate in an ADM process. A model satisfies the definition of no proxy discrimination if a protected attribute does not affect the decision via a proxy. It also satisfies no unresolved discrimination if this is the case. However, a protected attribute is allowed to affect the decision through what is called a *resolving* attribute. A resolving attribute is an attribute that is considered a legitimate and non-discriminatory variable (e.g. credit amount) that affects the prediction (e.g. credit score). (Verma & Rubin 2018, 6–7; Kilbertus et al. 2017.)

[66] Two approaches can be further distinguished. In a "human-on-the-loop" approach, the system scores and enacts decisions autonomously, although under the supervision of a human. The human may have an option to override the system's decision. A "human-in-command" approach subsumes the "human-in-the-loop" approach but extends human control even further, granting the human control over whether an ADS ought to be used at all in certain situations. (Citron & Pasquale 2014, 6–7; Ethics Guidelines for Trustworthy AI 2019, 16.) The distinction between out-of-the-loop and in-the-loop decisions will suffice in the context of this study, however.

of user bias are distinguished here: *transfer context bias* and *interpretation bias*. (Danks & London 2017.) In addition, the risk of feedback-loops and negative spirals is considered below.

2.4.1. *Transfer Context Bias*

Transfer context bias manifests when a model is deployed in a context that differs in some relevant manner from the model's intended context of use. If the standards, rules and affordances of the context deviate from those in the intended context of use, the ADS could exhibit perform in an undesirable or biased manner. An example of transfer context bias is provided by Danks & London who consider an autonomous vehicle intended for use in the U.S. and trained on U.S. traffic rules, respectively. "These autonomous systems", they say, "would clearly perform in a biased (in the negative sense) manner if they were deployed in, say, the United Kingdom, since people drive on the left-hand side of the road there" (Danks & London 2017, 4694).[67]

2.4.2. *Black Box Algorithms: Interpretation Bias and Opacity*

One source of user bias can be located in the semantic interpretation of the output. So-called interpretation bias consists in a deviation between what the output of an ADS actually represents and the semantic information a human user interprets the output to convey. For example, Danks & London note that humans (in-the-loop) may interpret "non-zero regression coefficients" as "indicating degree of causal strength" (2017, 4694). They see this as problematic, because even if there were a causal connection, a non-zero coefficient will not inform the direction of causality. In other words, an output may be interpreted as an indication of causality, perhaps even the direction of causality between variables, although regression coefficients only tell that there exists a negative or positive correlation between variables of some strength. (I consider the relationship between statistical and causal evidence in chapter 3, where I argue that the requirement for a causal connection as an epistemic justification for discrimination is problematic.)

Depending on the complexity of the model, specifying the exact semantic content of an algorithm's output may be more or less difficult. Algorithmic decisions may be based on thousands of variables and this will create difficulty in interpreting what exactly the output denotes. Danks & London note that an output value of a visual surveillance system (e.g. one using facial recognition) might afford different interpretations: When a subject is flagged as high-risk, this might indicate either

[67] Transfer context bias is distinct from training data bias in that while training data bias consists in the model's failure to generalize to the intended population, transfer context bias consists in the unwarranted or extended use of a model outside that intended context. (Danks & London 2017, 4694.) Notably, transfer context bias should also be distinguished from biased feature selection. These two are confused, for example, in Silva & Kenney (2018), who understand Danks & London's concept of transfer context bias to refer to bias that is transferred from other contexts into the model in the form of features or variables in the model. Danks & London, to my understanding, argue rather that transfer context bias results from the model itself being deployed in contexts beyond its intended context of use.

that the surveilled subject's identity is uncertain or that she is likely to engage in inappropriate activity.[68] (2017, 4694.) Whether it is the ADS or a human who enacts a decision due to the ADS flagging said individual as high-risk, it is arguably relevant whether the output value should be interpreted in one way or the other, as opaque outputs may result in unwarranted scrutiny (false positives) or neglected threats (false negatives). While the actual extent to which interpretation bias may occur in the use of ADS is uncertain, Danks & London state that "there is widespread potential for this kind of informational mismatch, since developers are rarely able to fully specify the exact semantic content (in all contexts) of their algorithms or models" (2017, 4694).

Interpretation bias is closely related to a paramount issue recognized in the literature on the ethics of AI: lack of transparency in algorithmic processing, i.e., the problem of "black box" algorithms (cf. Mittelstadt et al. 2016; Wachter et al. 2017; see also Jobin et al. 2019). As ML programs involve large amounts of complex code and extensive amounts of parameters upon which ADSs base their decisions, explanations for algorithmic decisions may prove difficult to extract. This is a problem with deep learning systems in particular because they comprise multiple layers of non-linear information processing. Tracking the exact logic behind the process where input is transformed into output in these systems may be a severely challenging task, even if one understands what the algorithm does at a general level. This problem of explaining the complex algorithmic processing can be referred to as the problem of model *explainability*. A related problem which is often distinguished from explainability relates to *interpretability*. The decision function of interpretable models (e.g., decision trees based on IF-THEN rules or linear regression models with two model variables) is often fairly easy to understand in the sense that a user can predict what the algorithm will do next or what will happen if the user intervenes on the input data (cf. Rudin 2019). In more complex models (e.g., multivariable models and deep ANNs), however, the decision function may be hard to understand even at a general level. Some have claimed due to complexity and high-dimensional data processing, the functions and outputs of some ADSs may be in some cases inherently uninterpretable in natural language (cf. Burrell 2016; Zarsky 2013, 1519). Danks & London's example above illustrates this problem: the decision-making process and the output of the visual surveillance system are uninterpretable if a user is not able to understand what the algorithm does even at a general level.

2.4.3. *Trust, Biased Assessment and Priming*

A human "in the loop" will introduce a mediating factor in how and which outputs are enacted on. This could potentially mitigate possible disparate impact or exacerbate it, irrespective of issues with

[68] A similar concern is expressed by Zarsky as well, who says that "[d]ata mining might point to individuals and events, indicating elevated risk, without telling us why they were selected" (2013, 1519).

interpretation. For example, humans in-the-loop could disregard their algorithmic informants and refuse to enact on every instance where an individual is flagged. In their case study on the AFST, Chouldechova et al. found that humans in-the-loop were "largely continuing to rely on their own assessments rather than those of the AFST tool" (2018, 12). Human-algorithm interaction can, however, also exacerbate disparate impact when compounded with unconscious cognitive biases humans exhibit (e.g., confirmation bias), and human decision-makers can exhibit selectivity in enacting those decisions. A controlled experimental study by Green and Chen (2019) showed that participants' assessments of defendants' risk for recidivism exhibited racial bias, even in relation to an ADS' preliminary assessment: "In every treatment, participants on average deviated positively (toward higher risk)" relative to the algorithm's output, that is, "for black defendants and negatively (toward lower risk) for white defendants" (Green & Chen 2019, 18). The central point here, then, is that human cognitive biases may compound with algorithmic outputs and thus affect what a given decision, or a distribution of multiple decisions, ultimately shapes up to be.

An associated risk relates to trust in ADSs. If a human decision-maker is inclined to perceive machines as more objective or trustworthy than humans, an ADS's output might override an opposing, possibly more nuanced judgment the decision-maker would have made in the absence of an algorithmic informant. Empirical studies on so-called *automation bias*[69] seem to support this notion. A survey study by Packin (2019) showed that people preferred an ADS's assessments as opposed to those of human experts in financial decision-making. Moreover, the study showed that people continue to do so even when an algorithmically generated decision was deemed disappointing. (Packin 2019, 21–23.) It has also been suspected that algorithmic outputs may psychologically prime decision-makers in ways that could lead to problematic consequences. For example, Selbst expresses the concern that when a predictive policing algorithm flags an individual or suspect as high-risk, officers handling the case "might proceed anxiously—with an itchy trigger finger—or otherwise be more easily provoked into unnecessary force" (2017, 137).

In any case, the introduction of human-algorithm interaction in real-life decision-making processes may have at least a potential effect of either mitigating or exacerbating possible disparate impact. The notion that fairness should be understood in this way, from a sociotechnical perspective, is emphasized also by Selbst et al. (2018). As I suggest below, that sociotechnical

[69] Dzindolet et al. (2003) found that people are prone to deem automated systems trustworthy even after limited interactions. However, perceived trustworthiness was found to quickly decline when the system made errors, although half as many as a human would make. For a review on the frequency and effects of automation bias see Goddard et al. 2011.

contingencies encompass the phenomenon of algorithmic discrimination in actual contexts is relevant from the point of view of moral theory.

2.4.4. *Feedback-Loops and Negative Spirals*

A prominent concern related to biased models in AD is that of negative feedback-loops. Training data bias may lead to feedback-loops where an oversampled part of the population will be subsequently targeted by an ADS. A discriminatory feedback-loop is discovered by Lum & Isaac (2016) in the use of the predictive policing algorithm PredPol. Their case study shows that sampling bias resulting from over-policing of certain neighborhoods subsequently led the algorithm to target those neighborhoods. This in turn resulted in further over-policing and disproportionate targeting of minorities[70]. A similar dynamic was found in the use of the AFST. As the system predicts re-referral (rather than child maltreatment *per se*) a negative feedback-loop ensues: When families are referred to the system, it raises their risk-score. This in turn makes it more probable that they are flagged in the future. (Eubanks 2018, ch. 4; see also Chouldechova et al. 2018.) Negative feedback-loops are especially concerning in that the scoring "dynamics will not be self-corrected, as they are misunderstood by the analysts studying the feedback of the scoring practices as mere reassurance of the scoring system's precision" (Zarsky 2014, 1405). That is, the predictions of an ADS may become self-fulfilling and validated as accurate in virtue of them (partly) causing the thing they are trying to predict.

Citron & Pasquale (2014) and Zarsky (2014) note the possibility that when individuals receive a negative score or outcome, that decision may send them down a "negative spiral". A negative decision may result from a "minor transgression" (e.g. one missed credit card payment) and yet lead to significant and, perhaps, disproportionately severe restrictions with respect to individuals' future actions (e.g. denials of future loan applications) (Zarsky 2014, 1406). Notably, negative spirals may ensue from both single- and cross-system feedback-loops: In situations of the former kind, previous decisions function as input data for new ones, for example. Cross-system feedback-loops may lead to negative spirals when one system's output (e.g., a negative decision) becomes the input of another system: A low credit score may negatively impact an individual's economic and social standing when that score is used in making employment decisions in another context, for example. (Citron & Pasquale 2014; Zarsky 2014, 1406–1407.)

[70] The risk of feedback-loops may be considered closely associated to bias in training data. Indeed, Lum & Isaac (2016, 15) note the following: "If police focus attention on certain ethnic groups and certain neighbourhoods, it is likely that police records will systematically over-represent those groups and neighbourhoods. That is, crimes that occur in locations frequented by police are more likely to appear in the database simply because that is where the police are patrolling."

2.5. Dissecting Discrimination in Algorithmic Decision-making

Having established different ways in which AD can have a disparate impact on groups, one can next distinguish generic characteristics of discrimination that may be attributed to algorithmic discrimination as well. These characteristics are often used in the legal literature to taxonomize distinct mechanisms of discrimination and to identify objectionable instances of discrimination in AD (cf. Zarsky 2014; Barocas & Selbst 2016; see chapter 2.2.1.). First, *sensitive classification* (or explicit discrimination) in the context of algorithmic decisions involves using (or representing) sensitive traits as features or class labels in a model, for example[71]. (Zarsky 2014, 1385–1386.) Sensitive classification requires that the model incorporates information that directly represents an individual's membership of a vulnerable group (i.e., sensitive information)[72].

However, as the examination of algorithmic bias shows, sensitive information may be also encoded in seemingly neutral data (i.e., data on unprotected attributes) as subtle patterns and values within the data. For example, COMPAS, a risk assessment tool used by U.S. Courts to predict defendants' risk for recidivism, was designed to predict recurring offences "within 2 years of assessment from 137 features about an individual and the individual's past criminal record" (Dressel & Farid 2018, 1). However, while an individual's race was not included in the 137 features that predicted the outcome score of a defendant, the outcomes were allegedly racially biased, nonetheless (Ibid.; see chapter 3.2. below). Indeed, an ADS may infer sensitive information from so-called redundant encodings in the data, and subsequently use it as a proxy in computing decisions. (Barocas & Selbst 2016, 691–692; see also Selbst 2017, 134.) Proxies for group-membership may consist in only one input variable or in a set of factors. In the USA, "[d]ue to housing segregation, neighborhood is a good proxy for race", as Barocas & Selbst note, but "[i]t is possible that some combination of musical tastes, [likes] on Facebook, and network of friends will reliably predict membership in protected classes" as well (2016, 712). Redundant encodings of protected group-membership (e.g. ZIP code) may, thus, be the result of past discriminatory treatment (e.g. housing segregation). Conversely, they may reflect individuals' self-representational behavior (e.g. likes on Facebook) that conform to behavior exhibited by other members of their group. I will call the use of redundantly encoded information about (vulnerable) group-membership in decision-making *proxy classification* (or proxy discrimination). It is important to note, however, that if a decision-maker were to engage in sensitive classification, this might prove trivial with respect to the outcome:

[71] Anti-discrimination law and ethical rules often prohibit the explicit use of sensitive information (or "formal classification") in decision-making processes (Barocas & Selbst 2016, 694–695; see also Zarsky 2014, 1385–1386.).
[72] Sensitive information may be erroneous, nonetheless, as it may mistakenly represent an individual to belong to a protected or vulnerable group while, in fact, she does not or belongs to another group.

> The only way using membership in the protected class as an explicit feature will change the outcome is if the information is otherwise not rich enough to detect such membership. Membership in the protected class will prove relevant to the exact extent it is already redundantly encoded. (Barocas & Selbst 2016, 695.)

In other words, the explicit use of sensitive variables (e.g., race) may be redundant if one's group-membership is encoded in other non-sensitive factors (e.g., ZIP code). Consequently, removing the sensitive variable may not mitigate model bias that follows from the use of such proxies.

In addition to sensitive and proxy classification, algorithmic discrimination is often categorized with respect to the role *intention* plays in using sensitive information. Disparate impact is most often an unintended outcome of AD because the algorithms used by data miners may inadvertently infer proxies for protected attributes (Barocas & Selbst 2016, 693). Models may inherit bias reflected by the training data, and decision-makers may fail to "de-bias" the model due to carelessness, negligence or ignorance. However, ill-intending decision-makers may also utilize the mechanisms examined in this chapter in order to intentionally discriminate against individuals and groups. For example, they "could knowingly and purposefully bias the collection of data to ensure that data mining suggests rules that are less favorable to members of protected classes" or they could coarse features in the model in an effort to produce higher rates of mispredictions for those members. (Barocas & Selbst 2016, 692) In addition, they could merely "preserve the known effects of prejudice in prior decision making", saying that the training data comprises "a reliable and impartial set of examples from which to induce a decision-making rule" (Ibid.; see also Zarsky 2014, 1389–1396.).

As intentional discrimination (e.g., intentional introduction of bias into datasets) may manifest only after an ADS has been deployed in use (e.g., as bias in the distribution of decisions for different groups) the problem of intentional discrimination is compounded by lack of transparency. While it could be prevented by certain countermeasures (e.g., audits and documentation of design processes), organizations may be reluctant to share details about the design and development of their ADS due to concerns about violating data subjects' privacy, disclosing trade secrets, or possibly enabling subjects to "game" the decision-making processes in their favor (cf. Wachter et al. 2017, 3–4). Moreover, decision-making processes executed by an algorithm could remain unexplainable and uninterpretable due to their complexity even if organizations were transparent with respect to technology design and use. Thus, multi-leveled opacity may enable organizations to conceal intentional discrimination from the public; it may effectively *mask* discriminatory intent. (Barocas & Selbst 2016, 692–694.)

Now, the possible combinations of these generic characteristics (intention and explicitness) exemplified by discrimination can be presented as follows (table 1):

Characteristics of discrimination	SENSITIVE CLASSIFICATION	PROXY CLASSIFICATION
INTENTIONAL	Sensitive information is intentionally and explicitly encoded in a model used for decision-making. (E.g., an individual's race is deliberately used as a predictor for creditworthiness.)	A proxy for sensitive group-membership is intentionally encoded in a model used for decision-making. (E.g., a proxy for an individual's race is deliberately used as a predictor for creditworthiness.)
UNINTENTIONAL	Sensitive information is unintentionally yet explicitly encoded in a model used for decision-making. (E.g., data miners accidentally fail to remove sensitive variables from the model before deployment.)	A proxy for sensitive group-membership is unintentionally encoded in a model used for decision-making. (E.g., data mining leads to a discovery of a proxy for an individual's race unbeknownst to the data miners.)

Table 1. *Interplay Between Intention and Forms of Classification in Algorithmic Decision-making.*

I suggest, however, that the occurrence of any dynamic presented in table 1. does not entail an *a priori* moral judgment regarding the permissibility of algorithmic discrimination. This is because they exemplify only characteristics of discrimination *in the non-moralized sense*. Whether there is deliberate aim to favor some group over another will tell us little about whether acting on that intention is justified. Similarly, that a model computes sensitive information does not tell us much about the outcomes of an AD process, or justification thereof. Even combined, the occurrence of none of these characteristics afford an *a priori* judgment of moral (im)permissibility.

Consider for example, intentional and sensitive classification. Zarsky states that "explicit and intentional discrimination is usually intrinsically immoral, as it features one individual incorrectly judging another to be of lesser moral worth" (2014, 1386).[73] It is correct that blatantly wrongful instances of, say, racist discrimination are often characterized by an intention to discriminate and feature a representational item that explicitly refers to race (recall the Montgomery bus policy). However, one should be skeptical regarding the claim that interplay between intentional partiality and explicitness would in all cases entail an intrinsic moral wrong. Differential treatment that draws on differences in sensitive traits can also serve morally praiseworthy purposes. As Andreas Mogensen notes, "[c]ertain affirmative action programs" in the field of education might exemplify this interplay

[73] Zarsky also notes that "explicit discrimination generates an additional specific harm given the actual and intentional use of membership of a protected class in the analysis process" (2014, 1386). In chapter 4, I will argue that this is partly correct, and may be so even when it is used implicitly and unintentionally. I will suggest that this additional harm is *expressive*; it is triggered in virtue of socio-historical contingencies, namely, histories of past injustice and systematic subordination.

in that they "may be thought to involve disadvantaging people on the basis of statistical generalizations about their race or ethnicity" (in the non-moralized sense), namely, non-minority members (2019, 456). But in doing so, it "may be expected to serve relevant educational goals such as breaking down stereotypes and promoting cross-racial understanding" (Ibid.). When vulnerable groups already suffer from disadvantage, favoring them in decision-making processes may serve to further an ideal that is not disrespectful of other groups. The aim may be to correct existing inequalities rather than to disadvantage some group.

In Eidelson's terms, when described "thickly", affirmative action can be morally justified even though it involves disfavoring some group in the process and the decision is based explicitly on differences in sensitive traits[74]. The relevant question, then, is not in the characteristic elements of discrimination – in the thinly described acts (i.e., racial classification or outcomes thereof as such) – but in "whether […] discrimination is permissible under [some] circumstances and for the sake of [some] ends" (Eidelson 2015, 62).[75] This point serves to only emphasize that discrimination is "a phenomenon with no built-in moral status" (Ibid., 14) although some characteristics of it (e.g., intentionality, sensitive classification, disparate impact etc.) are often used to identify objectionable instances of discrimination and, indeed, algorithmic discrimination. Of course, both the legal and public discourse understandably focus on identifying such characteristics because they *tend* to accompany a moral violation[76]. A philosophical account of algorithmic discrimination, however, should go further and identify every possible case of wrongful discrimination – i.e., meet Beeghly's identification condition (2017, 87; see chapter 1). Even in conjunction, intention and explicitness prove insufficient as means for such identification when in pursuit of a robust account of the ethics of discrimination. This argument will be provided further support throughout this study.

2.5.1. *Many Hands, Many Types of Bias*

Identifying wrongful instances of algorithmic discrimination seems like a difficult task, as the phenomenon in itself is multifaceted. This seems to be due to at least two problems. Firstly, the term "algorithmic bias" is used in several distinct senses, which creates confusion as to what sort of problem one is actually dealing with. It is used both to denote situations where an algorithm's "estimate [i.e., output] deviates from a statistical standard (e.g., the true population value)", but also to refer to cases where the systems output "deviates from a moral norm" (Danks & London 2017,

[74] Eidelson (forthcoming) explicitly defends the notion that affirmative action programs are not intrinsically disrespectful.
[75] Note that this suggestion also undermines the plausibility of defining fairness categorically as unawareness of individuals' sensitive traits (see chapter 3.2).
[76] Intentional and explicit discrimination, for example, could prove objectionable more often than not, at least in the eyes of the law. This is why explicit classification on the basis of, say, race could be considered a salient indicator of unlawful conduct, although it might be justified in some cases.

4692)[77]. In other words, an algorithm may be considered biased in a *statistical sense* when there exists a systematic distortion between the model (upon which it bases its predictions) and the population it is supposed to generalize to, or in a *moral sense* when there exists a systematic distortion between the way the algorithm functions as opposed to some normative demand or morally desirable outcome. These two senses are not mutually exclusive, however, as statistical bias can be introduced explicitly into the model in order to compensate for other instances of bias in the model, both statistical and "moral" – e.g., by mitigating data biases via using biased estimators (see chapter 2.3.1).[78].

Secondly, the tendency to identify wrongful "algorithmic" discrimination by looking only at whether AD has a disparate impact on some legally protected group – i.e., if it leads to unbalanced or "unfair" outcome distributions, such as COMPAS' allegedly racially biased predictions[79] – is problematic in two ways. Firstly, because data mining may lead to discovery of proxies that do not map out values of legally protected or sensitive traits, legal prohibitions on 'indirect' discrimination may not apply or be suitable to identify all instances of objectionable discrimination in AD (Mann & Matzner 2019; see chapter 2.5.3 below). Secondly, focusing on disparate impact seems to conceal some moral wrongs that could be identified at distinct stages of development and human supervised use of ADSs, at least in theory[80]. As Eidelson noted (see chapter 1.1.2.), 'indirect' discrimination is a coarse-grained concept from the point of view of moral theory in that it subsumes instances of both second-order (i.e., direct) discrimination and 'structural' discrimination, which are two distinct types of acts and may be wrong for different reasons. Second-order discrimination may be intrinsically objectionable due to basic disrespect, while 'structural' discrimination could be problematic not because it would constitute objectionable discrimination (recall that it does not satisfy the Differential Regard Condition), but because it maintains unequal social conditions, for example. In this sense, understanding discrimination in terms of outcomes (i.e., disparate impact) opposed to in terms of acts taking place in distinct dimensions of conduct creates difficulty in evaluating whether and when morally objectionable discrimination – or perhaps other violations of professional conduct – have occurred in the course of design, for example. Because it

[77] Danks & London also note that one can talk about bias in a legal sense, where the output deviates from some legal norm (2017, 4692).
[78] Similar to what is proposed here and in Danks & London (2017), Mitchell et al. (2018, 15) caution against confusing statistical, legal and ethical concepts of fairness and bias; they distinguish "statistical bias" stemming from problems such as measurement errors and unrepresentative sampling from "societal bias" which, for them, denotes (a retrospective notion of) social injustice.
[79] In the literature, there is a significant focus on the concept of disparate impact (cf. Barocas & Selbst 2016; Citron & Pasquale 2014; Zarsky 2014) As noted, this is understandable given the focus on legal concepts regarding discrimination.
[80] This is not to say that theorists have not paid attention to how choices in design may introduce bias into the models used for decision-making. Several theorists have considered issues with intentional biasing, confirmation bias and biased subjective deliberation (cf. Barocas & Selbst 2016; Zarsky 2014; Kaminski 2019).

may be that, on the one hand, the disparate impact is a result of *de facto* baseline differences between groups with respect to some attribute (e.g. criminal propensity). Then the relevant question will be whether it is *justified* that the outcomes of AD reflect those differences (i.e., whether they exhibit moral bias). On the other hand, it may be that disproportional outcomes are more or less explained by wrongful discrimination against some group at some stage of the design process – e.g., by interpretation bias in the use of an ADS.

Whether this is the case is hard to prove, however. As Mark Coeckelbergh eloborates,

> complex software often has a long history with many developers involved at various stages for various parts of the software. This can happen as software moves across organizations (e.g. a company) or even within the same organization. And in the case of machine learning AI there is also a process and history of the production, selection, and processing of data and datasets—again, with not just one human agent involved and happening at various times and places. There are […] people who collect and process data, people who sell data, people who analyze data, etc. AI software may also be developed in and for one context of application, but later used in an entirely different context of applications. (Coeckelbergh 2019, **7**.)

Indeed, part of the perplexity of the general notion of algorithmic discrimination seems to lie in the fact that there are multiple agents as well as several layers of decision-making and subjective deliberation at play. Consequently, statistical bias may be introduced in different ways, at different stages of development and use, and by multiple actors. In most cases, perhaps, this results in ill consequences. In some cases, however, it may also *mitigate* potential disparate impact (e.g. when models are refined to satisfy some fairness metric). In other words, different types of "bias", both statistical and moral, pile up. They interact but also counteract. Design choices and human involvement in use may comprise different "layers" of discrimination (in the non-moralized sense) in the process, which all causally contribute to how an ADS ultimately comes to "decide" who ought to be treated in what way – i.e., on what type of evidential basis it generates a decision. Determining the extent of the (causal) influence of a single agent, or of one design choice, on the bigger picture is difficult due to this "problem of many hands" (cf. Coeckelbergh 2019)[81].

2.5.2. *Distinguishing Types and Dimensions of Algorithmic Discrimination*

The notion of algorithmic discrimination, I suggest, can be clarified by analyzing it in terms of dimensions and types of discrimination, following Eidelson's conceptualization. Firstly, one ought to distinguish the process where an algorithm generates a decision as a distinct *dimension* of treatment,

[81] Of course, this is specifically a problem concerning responsibility-attribution and not discrimination *per se*.

conceptually (and temporally) separate from data collection, feature selection, and so on. It is an act of (using an algorithm when) generating a decision on the basis of statistical evidence in the dimension of decision-making. Secondly, that the decision is made based on statistical evidence is integral in terms of how one should conceptualize algorithmic discrimination. As has been shown, an ADS differentiates treatment (e.g. allocates goods or benefits) between individuals on the basis of the values a target variable receives. The (value of a) target variable itself functions as an evinced proxy, P, for some (value of a) target trait, T. Thereby, the Differential Regard and Statistical Evidence Conditions are satisfied (see chapter 1.1.4). (Whether a target variable is a *good* proxy, is a different question.) Furthermore, AD will *necessarily* involve comparative disadvantage for some group when there is some distribution in the decisions; i.e., for the group which receives an undesirable decision. Consequently, AD also falls under a distinct *type* of discrimination, namely, that of statistical discrimination[82]. It satisfies the Differential Treatment Condition, satisfaction of which is jointly explained by the satisfaction of the Differential Regard and Statistical Evidence Conditions. In effect, the model a decision-making algorithm relies on can be understood as a non-universal generalization (i.e., an algorithmic profile of sorts) and decisions made on the basis thereof will always include false negatives and false positives[83] (Zarsky 2014, 1409).

A crucial point I wish to emphasize here relates to the notion of *tainted* statistical discrimination touched on in chapter 1.4. Contingencies in data collection and other processes partly determine the outcomes of the treatment in the dimension of decision-making, in addition to whether there are actual baseline differences between groups that are captured by the model. But strictly speaking, these contingencies are located at distinct dimensions of treatment, and may involve *other types of discrimination* than that of statistical discrimination. Moreover, the data mining process and use of an ADS may also involve other moral issues *which are not inherently discriminatory* in that they concern every individual subjected to algorithmic decisions. If so, the statistical evidence used by an ADS may be tainted with disrespect, as it were. Thus, one finds a distinct set of moral questions regarding (i) tainted evidence and discrimination that follows from relying on that evidence, and another one concerning (ii) the justification of discrimination on the basis of statistical evidence altogether. I will next elaborate on this notion by analyzing instances of "algorithmic" discrimination in terms of distinct dimensions and types of discrimination in the development and use of ADSs.

[82] Indeed, as Barocas & Selbst note, data mining and AD are "*always* a form of statistical (and therefore seemingly rational) discrimination" (2016, 677). These notions apply to both supervised and unsupervised ML methods as the "rule" according to which an algorithm will differentiate between individuals is an approximated function of the data.
[83] Theoretically speaking, an ADS that would have a 100% accuracy would rely on a universal generalization. This seems technically implausible, however.

First, consider intentional algorithmic discrimination or *masking*, as Barocas & Selbst called it. If one follows Eidelson's conceptualization, any intentional decision to introduce statistical bias into the model in order to ultimately disadvantage members of some group should be conceived as an instance of second-order discrimination (see chapter 1.1.2.). To be considered morally objectionable, that discriminatory act will have be discriminatory *also* in the dimension of respect for personhood. Assume a malicious data miner intentionally introduces bias into the model at the stage of, say, data collection. If this is explained by the data miner's disregard for the equal standing of the discriminatees, it will be disrespectful and, thus, constitute objectionable discrimination. The fact that a set of training data is unrepresentative may be explained by the data miners' failure to consider each group's (e.g. ethnic minorities') equal interest for representation[84] and ultimately, accurate decisions. Thereby, it may be taken to violate the Interest Thesis (see chapter 1.3). Notably, the relevant dimension of treatment here (i.e., data collection and preparation), W_1, is separate from the dimension of decision-making, W_2, which involves *statistical* discrimination. The discriminatory act in W_1 of course contributes to means for decision-making in W_2 (i.e., the model), compromising the model's validity or accuracy, but nevertheless constitutes a distinct discriminatory act. Similarly, if overly coarse-grained features are used to intentionally disadvantage people of color – what Barocas & Selbst call "digital redlining"[85] (2016, 692) – that discriminatory act takes place in the process (i.e., dimension) of feature selection. These and other forms of (intentional) second-order discrimination can be considered objectionable if they are motivated by beliefs about some people's lesser worth, for instance. However, discriminatory intent (if understood merely as purposeful differential treatment) will not be sufficient for *morally objectionable* algorithmic discrimination. It may also serve a reasonable, redistributive purpose (e.g. mitigating group-inequities in education via AD).

Recall that statistical bias may be introduced at different stages of development due to unconscious bias as well. Defining class labels and labeling training examples may introduce bias when subjective deliberation is required from the data miners (chapters 2.2.2. and 2.3.2.). Prior (possibly prejudiced) beliefs about the phenomenon under consideration may be, in a sense,

[84] Supposedly, each group would have such an interest in cases where it is in that group's interests to be accurately classified by a given algorithm. Of course, one could presumably object to data collection and equal representation for other reasons, such as privacy concerns or concerns about what the system is used for in the first place (e.g. unwanted personalized advertising).

[85] 'Redlining' refers to a racially discriminatory practice that has historically enforced housing segregation. It has a particular history in the United States. In 1934 the Federal Housing Administration (FHA) introduced mortgage lending policies which involved the use of color-coded residential security mappings in determining which neighborhoods should be considered secure for investment. The coding system explicitly marked neighborhoods with a high percentage of black residents with the color red (hence the name), indicating ineligibility for backing by the FHA. Due to these redlining policies, the black population was disproportionately denied of mortgage loans. For a short history of redlining see Lockwood 2020.

embedded into the structure of the technology or model. These cases will fall under second-order discrimination as well, even though unintentional, if they are explained by differences in how a data miner (unconsciously) regards the relevant groups subjected to algorithmic decisions. Notably, as both conscious and unconscious second-order discrimination ultimately fall under the category of direct discrimination, when instances of AD involve such conduct, those instances may be considered *intrinsically* wrongful. Insofar as this is the case, unconscious second-order discrimination in AD could be understood as an instance of discrimination denoted by the BTV; "bad technology" may be the result of unconscious disregard for diversity, for example. These instances of second-order discrimination considered above are instances of tainted statistical discrimination and roughly analogical to the example of the (unconsciously) sexist manager of a factory (see chapter 1.1.2).

In addition, humans in-the-loop may also introduce another "layer" of discrimination in the use of an ADS (chapter 2.4.2.). Here, the discriminatory act is not temporally precedent to the process of generating the output itself, but subsequent instead. A decision-maker could assess an algorithm's outputs in a selective and biased manner due to (un)conscious prejudice, which was the case in Green and Chen's study (2019; see also Selbst et al. 2018, 62). This is discrimination in the dimension of making algorithm-informed decisions. However, if the outputs are enacted on with significant selectivity due to bias (racial or otherwise), such cases should be understood as instances of tainted statistical discrimination: the discriminator's cognitive bias affects the way in which he *enacts* on statistical evidence, inducing a distortion between the model's predictions and the distribution in the decisions. In this sense, human-in-the-loop AD can also involve *direct* discrimination. Direct discrimination in AD may range from subtle, biased enactments of inputs to blatant and oppressive practices. To elaborate on the subtle kind, assume an algorithm predicts that two demographic groups are equally likely to default on a loan, but the human decision-maker's assessments deviate slightly from that fact, and this leads to a slight increase in negative decisions for members for some group. If the disproportionality in the decision distribution is explained not by the statistical evidence but, rather, by the human agent's bias towards some group, the Joint Explanation Condition is not satisfied (see chapter 1.1.4.)[86] and the definition for direct discrimination will be met. An example of blatantly oppressive AD practices would be the Chinese government's deliberate and explicit use of facial recognition systems as a means for tracking and controlling Uighur muslims (cf. Mozur 2019). Again, these instances – blatantly oppressive AD practices in particular – should be

[86] Whether the racially biased assessments exhibited by participants in Green and Chen's study (2019) would fall under direct or statistical discrimination is unclear. Given that the deviations were – while statistically significant – rather minor, it would perhaps be problematic to say that the statistical evidence (i.e., the algorithm's assessment) had no effect on their decisions, although the biased assessments may be partly explained by prejudice or implicit cognitive bias.

considered as instances of tainted statistical discrimination: As Eidelson noted, "one *could* adopt a racial profiling policy by virtue of a failure to afford equal weight to people's interests" (2015, 179). Such a policy could even manifest contempt (see chapter 1.3) if the discriminatees' personhood *is* recognized and yet they are not treated as moral equals. Enacting on algorithmic outputs in an implicitly biased manner will also constitute tainted statistical discrimination (in this case wrongful direct discrimination) given that the Joint Explanation Condition is not satisfied.

Lastly, even in the absence of direct (second-order) discrimination, when AD has a disparate impact on some group, the development and use of ADSs will also involve 'structural discrimination' (see chapter 1.1.3.). Recall, that according to Eidelson's view, structural discrimination does not fall under the category of discriminatory acts *per se*. In this light, structural discrimination may occur in design and development when disproportionalities between groups with respect to some attribute travel along the development process and are ultimately modeled into the ADS, but this happens unmediated by an agent's regard of the subjects. Consider the following example: As was noted previously, data miners may rely on external datasets when collecting data. If these data sets are unrepresentative (e.g. with respect to ethnic minorities) and they are taken into use without further examination or inspection, such conduct could be understood as an instance of structural discrimination. Such conduct enables (or fails to mitigate) subsequent differences in (true) prediction rates and outcomes by way of negligence or mere carelessness. This may not be mediated by how different groups are perceived or valued, however. Data miners might also opt using insufficiently precise features due to lack of more precise data, or due to such data being more difficult to obtain[87]. If this is the case, there seems to be at least a potential justification for such structural discrimination, because the data miners might here, in fact, satisfy the Interest Thesis: perhaps, they do not disrespect the equality of the data subjects', but due to external circumstances and limitations, they are not able to use the most precise information or comprehensive data as would be desirable.

The notion of structural discrimination in the development of ADSs also highlights an important notion related to data and model accuracy. A failure to update data sets (cf. Selbst' example of erroneous data in police records in chapter 2.2.2.) could be understood as structural discrimination which does not necessarily manifest comparative disrespect. While use of erroneous data can be conceived as morally problematic in itself, the fact that the use of such data leads to adverse effects

[87] For example, a data miner could be interested in data concerning an individual's criminal history. While such more precise data may be unavailable, the use of race as a model feature may account for such information – although less accurately – due to disproportionalities in different racial groups' baselines with respect to criminal history. This is, of course, problematic because these baseline disproportionalities may be explained by historical, systematic discrimination. As such, incorporating overly coarse features could reproduce this systematic disadvantage. (Barocas & Selbst 2016, 690.)

on some group is explained by the fact that this group is over- or underrepresented in the (erroneous) data in the first place. Importantly, however, even those *not* adversely affected by algorithmic decisions will in such cases have an equal claim to be treated on the basis of accurate information, although they cannot claim they are discriminated against. If *everyone* is subjected to decisions on the basis of erroneous data, this may be wrong, but not necessarily wrongful discrimination. Such a structural discriminator will fail to respect *everyone's* interest to be treated on the basis of accurate information. Thus, the conduct can be considered disrespectful but not comparatively disrespectful, and thereby not an instance of objectionable discrimination.

The examination conducted above is integral from the point of view of moral theory: It gives us reason to argue that the malicious data miner who introduced bias into the model at the stage of data collection due to his racial bias, for example, has acted in an *intrinsically* wrongful manner. Because such conduct is wrong regardless of its outcomes, one can hold that the discriminatory act conducted by the data miner is morally wrong even if the model bias is mitigated at further stages of design (e.g. it is identified later in fairness evaluation by another data miner and the dataset is revised as a result). In other words, the distinctions offered here allow us to claim that disrespectful discrimination in the design processes of ADSs are wrong regardless of whether disparate impact actually follows from the use of algorithms. Agents who have deliberatively failed may be judged to have acted wrongfully both (i) in cases where they *contribute* to the disparate impact of AD as well as (ii) in cases where the effects of their acts (e.g., model bias) have been neutralized by other agents later in the design process.

2.5.3. *The "Nothing Personal" Argument and the Synthetic Groups Question*

A separate set of moral questions arises when one considers what I call *unalloyed algorithmic discrimination*. This relates to the view concerning structural discrimination (in the global sense) as distinguished by Barocas (2014) – namely, the Structural Discrimination View. The worry was that even if "data miners are extremely careful, they can still effect discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes" (Barocas & Selbst 2016, 675). In other words, even well-intending actors with accurate and reliable technology may end up using models that exhibit algorithmic bias of the moral sort. They may reproduce existing inequalities "reliably" by modeling morally biased yet accurate data – e.g., arrest statistics that indicate ethnic minorities as more prone to criminal activity due to historical over-policing. Additionally, the prevalence of criminality may be due to poor living conditions or lack of access to education or social services. In those cases, one may not find objectionable instances of discrimination in the development of the ADS, nor in how humans interpret its outputs and enact on them.

Unalloyed yet morally biased algorithmic discrimination is problematic as it leaves room for what one could call the *"Nothing Personal" Argument*[88]: Once the discriminators face resistance and claims of wrongful discrimination, they might defer to the fact that the distributions accurately reflected in the training data and, ultimately, in the outcomes of AD as "moral bias", are beyond their reasonable control. Thus, allegedly, this would render such statistical discrimination morally neutral (or even praiseworthy), and the discriminators ought to be viewed as victims of circumstance; they have meant "nothing personal". Yet it seems that there has to be a moral case against maintaining the subordinate position of vulnerable, historically disadvantaged groups through AD, even if this is done unintentionally and the practice is motivated by reasonable aims, such as ensuring security by preventing crime or avoiding the release of potential reoffenders to society.

In addition to the challenge posed by the Nothing Personal Argument, there is another challenge that relates to the notion of social salience. As noted above, AD constitutes a form of statistical discrimination; it will involve comparative disadvantage for some set of individuals, and consequently, discrimination against those individuals as a group *in the non-moralized sense*. At least, this is the case whenever everyone does not receive the same decision and even if it produces absolute benefits for all individuals. This notion follows from the conceptualization of discrimination I have employed: If discriminatory acts are not limited to those that concern socially salient groups, any individuals specified by an ADS (e.g., as high-risk) and subsequently denied benefits, goods, or otherwise disadvantaged, will be discriminated against (in the non-moralized sense). Notably, these individuals may not share an identical set of traits; they may have different values with respect to distinct predictor variables and yet receive the same decision. In addition, those traits may not necessarily align with socially salient categories (e.g., gender). In these cases, the targeted group has popped up from the data via mining, and the relevant set of predictor traits *P* may be seemingly trivial (e.g., one's chosen brand of mobile phone or location data). Its members constitute what Zarsky calls a "synthetic group" (2014, 1407), a socially non-salient, idiosyncratic mixture of identities specified by an algorithm because they share some statistically evinced propensity (i.e., value of a target variable). Insofar as this group is comparatively disadvantaged, and this fact is explained by them belonging to this group, such a case constitutes discrimination in the non-moralized sense; an instance of what Mann and Matzner (2019) call *emergent discrimination*. A prominent question is, then, what makes emergent algorithmic discrimination morally distinct from instances of algorithmic

[88] I have chosen to name the argument this way as a reference to Lippert-Rasmussen's (2007) article "Nothing Personal: On Statistical Discrimination" in which he considers the question of the wrongness of unalloyed statistical discrimination.

discrimination where it is a vulnerable or protected group that is discriminated against. Call this the *Synthetic Groups Question* (or, the SGQ).

2.6. Chapter Summary

In this chapter, I examined how AD may lead to disparate impact on different demographics due to bias, both statistical and moral. I analyzed the perplexity of algorithmic discrimination in terms of distinct acts that can be located in distinct dimensions of treatment – at least in theory, that is. A central point I have tried to convey is that paradigmatic instances of algorithmic discrimination will involve instances of 'structural' discrimination – and possibly even malicious second-order discrimination – at distinct stages of the development. Instances of human-in-the-loop AD, where statistical evidence (i.e., algorithmic outputs) plays no significant part in the explanation of the outcomes, may involve direct discrimination, ranging from subtly biased interpretations of outputs to blatantly discriminatory uses of ADSs. While I cannot offer a comprehensive account of all "pathways" to disparate impact in AD here, a fine-grained analysis is crucial for tackling objectionable discrimination. Distinct types of discrimination (and other contingently associated moral issues) may require distinct countermeasures, such as auditing or cognitive training to mitigate implicit bias.

In many cases, objectionable disparate impact in AD may not be exhaustively explained by the actual distribution of values with respect to some attribute (or a set of them) in a given population, which the model is supposed to capture as objectively as possible. Decisions made with regards to design that require subjective deliberation and are susceptible to implicit biasing by human actors may skew the model and consequently have adverse effects when the algorithm is put to work in actual decision-making contexts. The relevant dimensions of (discriminatory) treatment here are also often opaque which partly constitutes the perplexity of the phenomenon; those subjected to algorithmic discrimination may not know what happens "behind the curtains" – i.e., what choices were made in the development and use of ADSs. Furthermore, as I noted, AD may involve other moral wrongs, such as use of imprecise or faulty data. Negligence and carelessness, however, may be disrespectful of everyone while not being comparatively so. To constitute discrimination, the accuracy of the data should vary relative to data subjects' group-membership, respectively. My claim, then, is that not all moral issues with respect to AD amount to wrongful *discrimination* while they may be disrespectful all the same. Moreover, these issues may exacerbate and overlap with objectionable algorithmic discrimination that would surface regardless of those contingencies merely because the training data reflects inequality and injustice pertinent in society – and does so *accurately*.

Insofar as my examination has been sound, algorithmic discrimination – in the sense that the term is commonly used – can be realized through many distinct mechanisms or pathways of discrimination. It seems each of the views distinguished by Barocas (2014) correctly identifies one such mechanism in AD. Firstly, organizations and decision-makers can intentionally disadvantage groups by inferring sensitive information about them, i.e., their membership of a protected group (Bad Actor View). Intentional sensitive classification *per se* does not constitute wrongful discrimination, however; it can also serve morally praiseworthy purposes (e.g. affirmative action). Secondly, bad technology may result in disparate impact (Bad Technology View), but statistical bias itself does not exhaustively explain all wrongful instances of algorithmic discrimination. Statistically biased models may be even desirable – they may perform more accurately, and bias may serve as a means of regulating an ADS's actions according to some ethical principle (see chapter 2.3.1.). An example of the latter includes introducing statistical bias into the model (e.g. by removing problematic variables) for it to satisfy some fairness metric. The wrongness of instances denoted by the Structural Discrimination View, however, is yet to be explained. To do so, we have to answer the "Nothing Personal" Argument and the Synthetic Groups Question without reference to morally problematic contingencies in the design and human controlled use of ADSs. I will next consider two types of objections that would allegedly explain the wrongness of such *unalloyed* algorithmic discrimination.

3. Non-Contingent Objections Against Algorithmic Discrimination

In this chapter, I consider non-contingent objections against statistical discrimination. According to these distinct objections, statistical discrimination is intrinsically wrongful when it has a disparate impact on vulnerable or legally protected groups. I look at two lines of arguments or reasons as to why this would be the case, in particular. Allegedly, wrongful instances of algorithmic discrimination are wrong when and because (i) they involve a failure to treat people as individuals; or (ii) they are unfair in that they involve unequal treatment or partiality. I consider these objections in contexts where vulnerable or legally protected groups, in particular, are ones that are discriminated against. This is because similar objections have been expressed prominently in discussions concerning racial profiling (cf. Eidelson 2015, ch. 6.2.2.). If it can be shown that these objections do not hold *necessarily* even in those contexts, it is plausible that there is some other moral wrong underlying the phenomenon of algorithmic discrimination than the mere fact that sensitive traits play a part in it.

The first objection (i) states that all wrongful algorithmic discrimination is wrong because fails treat people as individuals. When an algorithm indicates an individual (e.g. a racialized person) as high-risk, unworthy of credit, or otherwise less deserving, the argument goes, her individuality is not shown adequate respect. Allegedly, this is because the generalization is inaccurate, the treatment is based on statistical as opposed to causal evidence, or she is objectified in that she is seen only through the lens of her reference-class Racial profiling, for example, is taken by some to be impermissible because it objectifies people of color or because correlations between race and crime (even if accurate) do not constitute individualized evidence (cf. Ibid.). According to the second objection (ii), racially or otherwise biased algorithmic discrimination fails to treat people as equals. Allegedly, it will violate the principle of fairness or equal treatment. Often underlying the objection from unfairness is the notion that discrimination is wrong when traits which an individual has not chosen or cannot change have affected the decision. As these objections translate quite seamlessly into contexts such as recidivism risk assessment, predictive policing, and credit-scoring, I will hereby use these practices as my primary examples.

To avoid repetition with issues considered in the previous chapter, I will focus on examining what I have called unalloyed algorithmic discrimination. In such cases, algorithmic decision-making processes, firstly, exhibit only some reasonable or justifiable amount of statistical bias, i.e., the predictions are reasonably accurate. (I will, however, consider also the relationship between accuracy and statistical discrimination in chapter 3.1.1.) Secondly, the algorithm's output is enacted on in an unbiased manner and thus issues with interpretation bias and priming are ruled out

of the picture[89]. Thirdly, the practices are meant to serve reasonable aims, such as preventing crime. Lastly, the decisions reproduce inequality between groups because the data accurately tracks existing inequalities and background injustice. If one or more of the objections stated above were to hold, algorithmic discrimination will be intrinsically wrongful even if it exemplifies none of the contingencies examined in the previous chapter, given that it adversely affects a protected group.

I suggest that algorithmic discrimination does not necessarily manifest the qualities ascribed by these arguments and is therefore not intrinsically wrongful as a form of statistical discrimination. Insofar as my argument holds, relying on sensitive information about individuals (either explicitly or via proxy) does not render algorithmic discrimination inherently wrong. A point of clarification is necessary here: Many (or most) real-life instances can involve contingent moral issues, such as inaccuracy or unfairness, considered throughout this study. I do not wish to claim otherwise. Rather, my aim is more limited in scope: I wish to examine whether the wrongness of *all* wrongful instances of algorithmic discrimination can be explained in terms of such issues. I suggest that they cannot. This conclusion conforms to views defended by Eidelson (2015) and Lippert-Rasmussen (2007; 2014), according to which (unalloyed) statistical discrimination is contingently wrong, if at all.

3.1. Failing to Treat People as Individuals

Algorithmic discrimination relies on probabilistic evidence about tendencies within different groups. According to the first set of arguments, an agent engaging in statistical discrimination fails to treat the subjects as individuals, and so acts in a morally objectionable manner. When an individual is considered only as a member of some group, the idea goes, her individuality is not afforded adequate respect. I consider three different interpretations of this argument, which are the following:

Discriminating against people on the basis of statistical generalizations (e.g., algorithmic profiles) is disrespectful of people's individuality because

(*The Inaccuracy Argument*)	they are inaccurate; or
(*The Causal Connection Argument*)	they rely on correlations as opposed to causal evidence; or
(*The Objectification Argument*)	people are seen only as members of a reference-class.

[89] Alternatively, one could rule out these contingent issues out by focusing on human-out-of-the-loop AD, e.g., credit-scoring or recidivism risk assessment where the ADS executes decisions autonomously.

I will next examine each of these objections. I conclude that neither inaccuracy, alleged objectification, or lack of information about causality exhaustively explain the wrongness of unalloyed algorithmic discrimination nor identify the right extension of cases which are wrong.

Two points of clarification are necessary: Firstly, a distinction can be made between treating people *individually* and treating them *as individuals* (Eidelson 2015, 136). In AD, the problem is likely not that people would not be treated individually because (most) decision-making processes do not deal with groups, but individual persons (e.g., loan applications are evaluated individually). Secondly, if an ADS would consider *only* differences with respect to some sensitive trait (e.g. a model would incorporate only one variable, such as gender) and this would lead to discrimination, it would be straightforwardly wrong; it would blatantly fail to disrespect the individuality of the discriminatees (see chapter 1.4). However, this is not a necessary quality of profiling, and even less so with algorithmic profiling. Algorithmic discrimination, as I have shown, most often involves multiple proxies that may, however, map values of sensitive traits to a varying degree (e.g., recall that COMPAS uses over 130 factors).

3.1.1. *The Inaccuracy Argument*

The Inaccuracy Argument has been invoked in the context of discrimination conducted on the basis of stereotypes (cf. Beeghly 2018): stereotyping and profiling are allegedly wrong because they involve insufficiently accurate means for differentiating treatment (i.e., non-universal generalizations). But are algorithmic profiles extracted from data via data mining techniques necessarily disrespectful of persons' individuality *because* they are inaccurate? Arguably not. While discrimination on the basis of generalizations is often objected to due to worries of inaccuracy, "inaccuracy cannot be the core problem" of generalization (Beeghly 2018, 691; see also Binns 2017 and Lippert-Rasmussen 2007). Statistical generalizations may in fact be highly accurate, and when they are, relying on them in differentiating treatment is rarely objected to. Consider DNA evidence, for example: While DNA evidence is statistical in the sense that it "identifies the frequency with which genetic profiles occur in reference to populations", it is rarely held insufficient as evidence for conviction (Enoch & Fisher 2015, 587).[90] Convictions have been made in U.S. Courts on the basis of DNA evidence that corroborates other evidence, but also on the basis of DNA evidence alone. DNA

[90] Another prominent example is functional magnetic resonance imaging, or fMRI. Images produced via fMRI are used in the field of medicine to differentiate between those who should get medical treatment (or who require medical intervention) and those who do not (at least aside other evidence). Peculiarly, however, fMRI images are not images in the same sense as photographs, for instance. They are, rather, inferences about brain activity based on blood flow and oxygenation patterns in the brain which are then represented in visual form. In this sense, they are visual models that represent brain activity via blood flow. For an introduction to functional magnetic resonance imaging, see Jezzard et al. 2001.

evidence has even triumphed over a conflicting witness testimony in some cases. The value of DNA evidence may be partly explained by its rather low margin of error (i.e., small likelihood of false positives). (Ibid., 588–589.)

The problem with the Inaccuracy Argument is that, even if some statistical generalizations were inaccurate and thereby less valuable in practice, this does not point to a general moral flaw in the use of statistical evidence altogether. It only entails that less accurate statistical evidence is less preferable. It seems, then, that "criticisms of generalization in general […] may in fact boil down to criticisms of *insufficiently precise means* of generalization" (Binns 2017, 5). But moreover, when we consider blatant stereotypes (e.g. racial generics), it seems that even if they *were* accurate, stereotyping people could still be regarded as morally problematic. Associating racial categories to problematic traits (e.g., criminality) may violate respect-conventions and produce harm and feelings of denigration. Likewise, as I have suggested, it seems plausible that wrongful discrimination may ensue even when algorithms operate on accurate data. Even if, say, arrest data were accurate, it is plausible that it represents a symptom of underlying patterns of discrimination, background injustice and social conditions that lead to increases in discovered crimes. This was what the Structural Discrimination View (SVD) posited: AD may lead to objectionable reproduction of systemic inequality. Arguably, then, while inaccuracy is a problem in itself, it does not constitute an independent, underlying moral wrong of statistical discrimination.

3.1.2. *The Causal Connection Argument*

The second objection, the Causal Connection Argument, states that statistical evidence does not warrant differential treatment of individuals because it is not *individualized evidence*; evidence that "is causally linked in an appropriate way to the thing for which it is taken as evidence" (Enoch & Fisher 2015, 567). By contrast, treating individuals differently on grounds of causally connected evidence (e.g. detaining a racialized person on the basis of a witness testimony) the treatment would be justified in that there will be an appropriate causal link between an individual and the target trait in question (e.g. crime). Mittelstadt et al. (2016) have expressed concern over the fact that algorithmic decisions are grounded in correlations rather than information about causality. The worry is that AD may disadvantage individuals not because of what they *have* done but, rather, because of what they are *likely* to do in light of patterns in data.[91] Supposedly, then, sensitive attributes should not contribute to an ADS's output, as they are often not causally connected to, say, one's creditworthiness or risk of reoffending.

[91] This particular concern is also referred to as "arbitrariness-by-algorithms" (cf. Citron & Pasquale 2014; Zarsky 2014).

The Causal Connection Argument has four variants which can be considered in the context of racial profiling. Indeed, racial profiling is taken by some to be intrinsically wrongful in this regard; an individual's race is not causally connected to criminality, even if statistical evidence shows that the two co-occur (cf. Eidelson 2015, 192–196; Enoch & Fisher 2015, 567.). It proves a fruitful example here because the arguments against racial profiling translate seamlessly into the context of predictive policing, if one takes the relevant statistical information to be inferred via data mining[92]. These variants, each concerning a distinct direction of causality (or lack thereof), can be presented as follows:

Racial profiling (as a form of statistical discrimination) is intrinsically wrong because

(i) race is never downstream evidence of crime, and only such evidence justifies scrutiny; or
(ii) race is never upstream evidence of crime, and only such evidence justifies scrutiny; or
(iii) race is neither down- or upstream evidence of crime, and only those justify scrutiny; or
(iv) crime and race bear no causal or explanatory connection with each other, and there must be one to justify scrutiny.

The first objection implies that while a person's being a shooter would explain a gun found at his apartment, it will not explain an individual's being of a certain race. Thus, singling out people on the basis of their race will be unwarranted even if there is a correlation between criminality and race. Conversely, the second objection would entail that because race does not explain the property of being a shooter, information about one's race is insufficient to justify scrutiny. (Eidelson 2015, 193.)

As explicated by Eidelson, the relevant question here is "whether or not an explanatory connection is required" in order to justify something as evidence, "does the presence or absence of such a connection bear on the morality of subjecting someone to scrutiny on the basis of a particular trait?" (2015, 193). Eidelson argues that the answer is no: the objections are mistaken in that they rely on an unwarranted notion according to which race (or any other sensitive trait) would bear greater *inherent* moral significance in comparison to many other traits or things that are used as evidence (for crime). Both up- and downstream evidence are regularly deemed as justified triggers for scrutiny, he argues, and reliance on race does not differ from those in a way that would make profiling *intrinsically* wrongful. (Eidelson 2015, 193–194.) He maintains that objecting to the use of information about an

[92] As noted above, most often other variables, or aggregates thereof, function as *proxies* for race. However, if it can be argued that even explicit use of race in decision-making can be justified, then it seems that the use of proxies can be as well.

individual's race as (both up- and downstream) evidence would mean that we could not rely on other sort of evidence albeit they are commonly deemed as genuine evidence, and justifiably so.

The first objection is problematic, according to Eidelson, as it would exclude relying on (evidence of) motives as a part of ordinary police work, for example:

> By definition, motives explain crimes, rather than being explained by them. Nor is *evidence* of a person's motive normally explained by his committing a crime. But it seems to be perfectly genuine evidence, and it is often admissible in criminal trials. The standard for imposing searches, one would think, should be no more stringent. (Eidelson 2015, 193–194.)

In other words, if justifying scrutiny requires that the target trait (e.g. committing a crime) explains the indicator trait on the basis of which one is scrutinized (e.g. race), we would have to refrain from using entirely common and seemingly legitimate non-downstream evidence altogether (e.g. evidence of motives).[93] Notably, if we were to object to the use of non-upstream evidence, this would be problematic all the same. A bloody murder will explain blood stains on a carpet, for example, and it would seem rather absurd if this kind of downstream evidence were to not justify scrutiny. Thus, the second objection seems to face a similar problem as well. (Eidelson 2015, 194.) What about the third variant then? According to the objection, if something is neither down- nor upstream evidence of crime, it will be objectionable to single out people on basis of that evidence. Eidelson notes that when there are correlations "between demographic variables and crime patterns", we should at least assume that these "are explainable" at least in principle, no matter how complex the patterns are (2015, 194). One possibility is that the correlation is explained by a common cause; some unknown variable should explain why crime and race are connected. Plausibly, prevalence of crime in this case may explained by poverty and other social factors, including background injustice; past and present discrimination and oppression suffered by racialized groups (cf. Risse & Zeckhauser 2004). Regardless, if we take (iii) to be correct, we should concede that it is impermissible to use evidence that ties two things together only via a common cause.

Disregarding such evidence in investigation seems problematic, however. Eidelson notes that such evidence "can easily be constructed simply by introducing an observable consequence of something that would otherwise serve as upstream evidence" (2015, 195). Consider the following example[94]. Mary and Matt are recognized philosophers, widely known for their mutual passion for Plato's work. They have long planned a trip to a conference revolving around the Greek philosopher's

[93] Furthermore, it would be implausible that motives would causally explain (at least by themselves) a specific crime. People may have strong motives for conducting crimes, but not all who have such motives enact on them. In this sense, motives rather correlate with crime than exhaustively explain them.
[94] This is a slightly modified yet analogous example to what Eidelson (2015, 195) presents.

writings. Suppose that Mary, despite their prior commitment, confesses to Matt that she is done with Plato's philosophy, and moved on to that of Aristotle's. Hence, she will not be attending the conference after all. Matt, enraged by this betrayal, kills Mary with a copy of Plato's *Republic*. Suppose now that police apprehend Matt when they find a blood-stained copy of the Greek classic in his office. Matt does not mention Mary's confession to the police, however. A week later, the police happen to find that Mary had described her heated confrontation with Matt in her diary at the night of the murder. Now, the diary-entry will be explained by Mary's confession, as will be the killing. Mary's confession is, thus, a common cause of both. The diary-entry in itself, however, does not explain the killing. Yet the "difference *in the explanatory structure* does not seem to yield any significant *moral* difference" for whether Matt is treated fairly when the police rely on either the confession itself or the diary-entry *as evidence* for crime (Eidelson 2015, 195; italics added). Thus, Eidelson maintains that one should doubt the idea that reliance on common cause evidence should be objectionable.

Should racial profiling be inherently wrongful on the grounds that there is no explanatory connection at all between race and crime, as the fourth variant states? Claiming that "certain kinds of crime really are not probabilistically independent of race" seems contrived, Eidelson argues; "there should be a strong presumption that *some* explanatory connection, however attenuated, unites the two" (2015, 195). Denying the "empirical reality" that there are genuine correlations between sensitive traits and criminality, he notes, "is a losing strategy for critics of racial profiling" (Ibid., 175). This is because even if the evidence reflects only past discrimination and impermissible conduct, questioning the existence of that evidence (as opposed to what it actually reflects or how those statistical facts have come to be) does not constitute a fruitful strategy for combating racial profiling. Notably, this does not by any means entail that one should concede claims about certain racial groups' having an inherent disposition for crime.

For Eidelson, it seems that even if we supposed that there is no causal connection between race and criminality, the existence of a (strong enough) correlation between those traits will warrant scrutiny on those grounds. His argument is grounded in Bayesian epistemology, or more specifically, in Bayesian confirmation theory. A core notion of this theory is that any

> evidence [E] confirms (or would confirm) hypothesis H (to at least some degree) just in case the prior probability of H conditional on E is greater than the prior unconditional probability of H: $P_i(H|E) > P_i(H)$. E disconfirms (or would disconfirm) H if the prior probability of H conditional on E is less than the prior unconditional probability of H. (Talbott 2016, S.4.2.)

The theory entails that when some statistical evidence (i.e., an indicated probability) is stronger, there is a stronger warrant for an increase *in the degree of belief* regarding some fact. Evidence to the fact that *Y* has some predictor trait *P* serves to confirm the hypothesis that *Y* is likely to possess the target trait *T* (e.g. committing crime), insofar as we have evidence that *P* predicts *T* to a reasonable degree. In other words, if we find that *Y* is a member of a reference-class specified by *P*, this will function as an epistemic warrant for targeting that individual, insofar as the evidence that *P* co-occurs with *T* is sufficiently accurate and one has no disconfirming, stronger evidence. (Eidelson 2015, 183–185.) More importantly, however, it does not make an *a priori* difference as to whether *P* is sensitive (or not) with respect to the morality of using knowledge thereof as evidence. As far as the epistemic justification of such inferences goes (which is under consideration here), using a racial profile does not differ in that regard from other common practices which involve statistical discrimination. These include searching for possible disease carriers by way of employing medical profiles or attempting to find drug smugglers by scrutinizing "passengers who pay in cash, check no hold luggage, and make frequent short-stay trips to key drug source cities" (Mogensen 2019, 456).

Furthermore, while so-called individualized evidence is often preferred to statistical evidence, there is a problem with grounding a distinction between the two by way of the causal connection view. Consider the use of suspect descriptions or witness testimonies that refer to a suspect's race (e.g. 'a member of ethnic group *Y* was seen at the crime scene'). Both are used as relevant evidence for scrutiny in police work. A problem comes from contrasting individualized evidence with statistical evidence in these cases, because the distinction seems contrived to begin with. What we take to constitute evidence of the former kind usually turns out to rely on generalizations and probabilities that we are merely accustomed to taking as granted or think of as being more reliable. Reliance on an eye-witness testimony itself involves a generalization that a witness' perception is reliable evidence, for example; a generalization which is not in itself waterproof[95]. The distinction is, arguably, "superficial because any non-statistical information qualifies as such only because it stands on the massive shoulders of statistical information" (Lippert-Rasmussen 2011, 51) In addition, Enoch & Fisher correctly point out that "courts may sometimes need to accept evidence (expert witness testimonies, for example) regarding certain mathematical truths", and it is difficult "to see how the causal requirement can be met here, given that mathematical truths are, arguably, causally inert" (2015, 567). For sake of consistency, then, objecting to the use

[95] As Schauer notes, the reliability of eye-witness testimonies is rather unsupported by empirical evidence. Eye-witness testimonies are influenced by problems with memory, cognitive biases, as well as eye-witnesses' beliefs about others' expectations regarding their testimony (2006, 94).

statistical evidence would thus require omitting commonly used (forms of) evidence that lack appropriate causal links in other practices, e.g., court rulings and conviction, as well.[96]

3.1.3. *The Objectification Argument*

The last objection, the Objectification Argument, is expressed by Margot Kaminski as follows:

> Automatically making decisions based on what categories an individual falls into—that is, what correlations can be shown between an individual and others—can *fail to treat that individual as an individual*. It makes her fungible, exchangeable with others, through objectification. If algorithmic decision-making does not allow an individual to *proclaim her individuality* […] then it *violates her dignity* and objectifies her as her traits, *rather than treating her as a whole person*. (Kaminski 2019, 1542; italics added.)

Binns entertains this objection as well: If the wrongness of algorithmic discrimination is explained by the notion that individuals are seen only as members of their reference-class, this "presents a strong challenge to the very existence of algorithmic decision-making systems; since they fail to treat people as individuals by design" (2017, 4).

Recall that, according to Eidelson, an agent will fail to respect a person's individuality by violating either the Character Condition or the Agency Condition, or both. (2015, 144.) The question is, then, whether AD inherently violates these conditions. Starting with the Character Condition, it was required that an agent considers reasonably available information that reflects how a person has exercised her autonomy. In the context of AD, this requirement translates to one specifying the kind of *data* that should be used in forming decisions. Kaminski's worry is that algorithms may fail "to consider individualizing factors" and produce "unfair outcomes" by representing "decisional subjects in too-broad strokes" (2019, 1543). While unfair outcomes pose a distinct concern (see chapter 3.2. below), I have hinted towards the notion that relying on statistical evidence in discriminatory practices does not necessarily entail disrespect for one's individuality. Indeed, Eidelson holds that the use of group-level statistical information in itself is not morally impermissible because it is "genuine information about its members" and "there is nothing wrong per se with making use of such information" (Eidelson 2015, 145). Rather, individualizing factors should

[96] One could also object to the causal connection argument on the grounds that it fails to account for cases involving "multiple causes, independent causal chains, different facts that suffice causally only together, different facts each of which suffices causally alone, etc." (Enoch & Fisher 2015, 567). Thus, it is not only too strict and has "intuitively unacceptable" implications, but it is also "not clear what follows from a causal theory and where" (Ibid.). Independently of the issue of statistical evidence, then, the causal connection argument may fall due to its inherent problems. For an examination of complex cases of causality in the context of law, see Moore (2019).

be considered *in addition* to this information; respect for subjects' individuality is essentially a requirement for the *inclusion* (as opposed to exclusion) of information in deliberation.

Supposedly, these individualizing factors are sufficiently considered, then, to the extent that the gathered data accurately reflects (i) information concerning an individual's actions; and (ii) information about the individual that may either set her apart from or affirm her conformity to the statistical tendencies associated with a given group she belongs to. It should reflect instances of self-representational behavior that a person has engaged in.[97] Were the decision-making process not to account for such factors, it will disrespect the individuality of those subjected to the decision-making process. Kaminski argues that there is a risk that both (i) and (ii) go unrecognized when engaging in AD. In relying on ADSs in deliberation and decision-making, she argues, "we potentially eliminate important work that a human decision-maker does to both fill in and circumscribe decisional context in a particular case" (2019, 1546). An ADS may fail to recognize relevant intentions behind subjects' actions or context-specific justifications for those actions. Kaminski uses the example of an individual driving over the speed-limit on her way to the hospital; an ADS might deem this worth penalizing while a human could consider speeding justified due to contextual factors, such as an imminent threat to human life (Ibid.). She also notes that ADSs may base decisions on rather arbitrary (albeit accurate) correlations – e.g., "connecting loan decisions to [one's] smartphone choice" (Ibid., 1547) – a mistake that humans equipped with adequate contextual knowledge would be less likely to do.

The Objectification Argument has some pull. Even though ADSs can plausibly account for *more* information in decision-making than their human counterparts (e.g. they can calculate a wider set of variables and be more consistent in their predictions) they seem to lack the capacity to consider possibly relevant contextual factors. These factors, e.g., individuals' intentions and mitigating circumstances, would plausibly constitute the kind of individualized evidence Eidelson is referring to. As I understand, the problem in question seems to be that algorithms lack what one could call *deliberative flexibility*: human decision-makers can choose to extend the scope of relevant information (i.e., factors) to be taken under consideration when necessary. As Eidelson states, respect in decision-making requires that we "regulate our conduct" in such ways that we come to appreciate the morally salient aspects of individual persons (2013, 210). Algorithms, however, calculate only a predetermined, fixed set of factors in each instance on the basis of a set of data they are provided with. As such, they are incapable of deliberative flexibility; a context-sensitive and adaptive

[97] Depending on the context of the decision-making, the nature of this information will vary, of course. For example, an individual can manifest her autonomous capacity by selecting products of interest to her while shopping online. However, omitting from clicking on ads (that may in fact interest her) on a social media platform, for example, could perhaps be regarded as an exercise of autonomy all the same.

regulation of decision-making processes. AD makes it unable to relate to the subjects whom decision-makers govern and, thus, it fails to capture individuals' unique *character*.

What about the Agency Condition, then? Recall that regarding a person and her future actions as "determined by statistical tendencies" is a violation of this condition (Eidelson 2015, 148). I argue there is a moral problem with AD in this regard, and it relates to the historic nature of data as such. As Kelleher & Tierney state, "data are representations of observations that were made in the past" and yet data mining and ML algorithms are used to "search through the past for patterns that might generalize to the future" (2018, 150). If so, relying on predictive models in decision-making will manifest disrespect for the subjects' autonomy-as-capacity. Treating an individual on the basis of her "data double" – that is, "a shadow self consisting of data points gathered about an individual" (Kaminski 2019, 1543) – ignores her capacity to author her own life; to exercise her autonomy by making reflective decisions that may or may not conform to those she has made in the past. Given that algorithmic systems predict future states-of-affairs from past data, it is difficult to conceive how ADSs could appreciate the fact that people are autonomous even in principle. Indeed, as Binns notes, given the deterministic nature of algorithmic systems[98], for "any two individuals with the same attributes" an ADS is bound to produce "the same output" (2017, 4–5). This seems to go against the requirement that we confront each individual *as* a unique individual, even if they are identical with respect to the set of factors that we have preliminarily decided to consider in making decisions. Due the way ADSs operate, it is unclear how people subjected to AD processes were not treated as if they were "determined by demographic categories or other matters of statistical fate", to use Eidelson's terms (2013, 216). That is all algorithms have available *by design*, as it were.

Nevertheless, one could counter this point: generalization seems to be an inevitable aspect of how *humans* make decisions as well. It "is something we can hardly avoid engaging in given that inductive reasoning and the tendency to act thereupon are deeply ingrained in our nature" (Lippert-Rasmussen 2014, 272; see also Schauer 2006). Eidelson echoes this line of thought by noting that the distinction between generalizations and other forms of evidence is more of a line drawn in the sand (2015, 147). As noted above, for him, respecting people's individuality is a question of inclusion of more information related to an individual in a given deliberative process. It is not a requirement to exclude some things from consideration, such as information about an individual's

[98] Artificial neural networks in particular are praised for their capacity to avoid catastrophic breakdown and indeterminate states, making them robust in that they can "find near-optimal solutions that satisfy multiple […] constraints" (Sun 2014, 110). That is, they are able to execute decisions deterministically even in sub-optimally informed decision-making contexts.

reference-class. It has been established that the nature of that information (e.g., whether it is statistical or causal) seems to make no intrinsic difference with regard to respectful decision-making.

Even if we were to concede the possibility that some instances of AD may be disrespectful of subjects' autonomy, there may yet be instances where AD does not involve basic disrespect. As Beeghly notes, Eidelson's formulation of the Character Condition entails "that there is nothing wrong with stereotyping [i.e., profiling] people when personalized information is irrelevant or not reasonably available. So not all stereotyping would violate the character condition" (2018, 702). Insofar as the relevant sort of data are not available, "employing gross statistical categories or blanket policies" may not be disrespectful (Eidelson 2013, 224)[99]. Indeed, recall that data may be unavailable due to several reasons, such as dark zones in data (see chapter 2.2.1.). If the data miners have *tried* to secure the relevant data – yet fail to do so due to some reason – they might not be understood as having violated the Character Condition.[100] (Note that such cases could constitute 'structural discrimination' in designing and training ADSs.) In these cases, it seems that one cannot claim that data miners have failed to respect the data subjects' individuality, given that there has been an adequate effort to obtain all relevant information.[101] As noted, it may still be wrong to make high-stakes decisions on the basis of insufficient or unreliable data. However, this means that when the relevant sort of data are not available, algorithmic decisions that have a disparate impact on protected groups are not *intrinsically* wrongful, at least on the ground that they would disrespect people's individuality.

3.1.4. *On the Alignment of Objectification and Discrimination*

There is another counterargument against the notion that algorithmic discrimination would be wrong in each instance because it fails to respect people's individuality. It draws from an integral distinction explicated by Lippert-Rasmussen:

> Discrimination is an essentially comparative notion: it requires that I treat different groups of people differently. But failing to treat someone as an individual is non-comparative. Hence, I cannot discriminate against everyone, but I can fail to treat everyone as individuals. (2011, 56.)

---

[99] Eidelson notes that such instances could be wrong "on grounds of fairness or reliability" (2013, 224). Coarse statistical generalizations may be inaccurate or rely on spurious correlations and thus constitute unjust treatment, but these are different problems with regard to their respective moral violations, nonetheless.

[100] Notably, what would constitute "reasonable availability" of information (or data) is, however, underspecified by Eidelson. Therefore, this part of the account remains somewhat vague.

[101] It could be argued that it is not self-evident that treating people on the basis of their individual traits and actions is always preferable to treating them on the basis of statistical information to begin with. As Lippert-Rasmussen notes, if we maintain the notion "that we ought to treat people on the basis of properties they in fact have", this entails that "we should favour practices ensuring that we do so to the highest degree" (2007, 395). Given that statistical information may in many cases be in fact highly accurate, there may be room to argue that we should favour it over individualized evidence in some cases.

If we consider the nature of algorithmic profiling, we see that AD could in fact fail to respect some people's individuality yet not constitute wrongful discrimination. Firstly, given that the generalizations (i.e., the statistical model) function as a basis for decisions across *all groups*, we might find cases where the set of people who are (wrongfully) discriminated against is not identical to the set of people whose individuality is not being afforded adequate respect. To elaborate, consider the COMPAS tool and its predictions. COMPAS will equally reduce every subject to a risk score on the basis of the set of factors it calculates. Regardless of *what* prediction it produces (e.g., low-risk or high-risk), it may or may not account for information that is required to satisfy the demand for respect for individuality with respect to each defendant. A defendant could benefit from the set of statistical rules it follows even if it were disrespectful of her individuality. Conversely, a defendant's individuality could be respected when a model captures all normatively significant factors with respect to her character, but she might still be discriminated against in the non-moralized sense.

Secondly, recall that Lippert-Rasmussen's (2019) critique of Eidelson's account of autonomy provided reason to supplement the account with a stricter comparative element, which I called the Equal Autonomy Condition. According to this condition, an agent *X* will wrongfully discriminate against *Y* by affording *comparatively less* normative weight to how she has exercised her autonomy or will do so in future than is afforded to some (counterfactual) other, *Z* (even if *X* in fact satisfies both the Character and Agency Conditions with respect to both *Y* and *Z* separately). It is plausible that decisions made on the basis of a statistical beliefs may even fail to disrespect *everyone's* individuality in some cases and yet satisfy the Equal Autonomy Condition[102]; each individual's individuality may go unrecognized, but to the same degree. Of course, it is desirable that statistical models account for heterogeneity and variance but, given that everyone may be objectified to the same degree (in principle, at least), AD does not necessarily violate the Equal Autonomy Condition[103].

I have suggested that lack of reasonable access to data may serve to refute the allegation that AD would necessarily involve disrespect for individuality. Moreover, I argued that even if AD were disrespectful of people's individuality, it will not in all cases involve discrimination in the dimension of respect for personhood (i.e., comparative disrespect for individuality). The underlying problem, as I will elaborate in chapter 4, might pertain more to what a given generalization *expresses*. Even if COMPAS' predictions were to portray individuals only as artefacts of statistical fate, only

[102] Age limits for voting, for example, depend on a very rough proxy for eligibility (i.e., age) even though it will not capture many relevant pieces of information regarding an individual's capacity to make informed decisions. It is hard to see how this policy would afford each individual's unique character its due normative weight. Furthermore, it may be unreasonable to expect that decision-makers could acquire such information. Indeed, statistical blanket policies do not necessarily even aim to be respectful of individuality but serve to meet population-level goals.

[103] An upshot of this view is that one could even remain agnostic as to whether AD inherently involves disrespect for individuality and still maintain that it does not amount to wrongful discrimination.

some people of these people are in fact deemed more likely *to re-commit crime* (which is an undesirable trait if any). Yet, statistical discriminators "are not expressing any belief that he or she has always been bound to do so" and neither are they "implying that he or she is anything less than a fully autonomous individual" (Enoch & Fisher 2015, 569). Noting the apparent tension, Eidelson says that the wrongness of racial profiling does not lie in the epistemological status of the inferences involved – i.e., in a lack of epistemic conscientiousness. Rather, it may have more to do with the social meaning that drawing certain inferences bears (e.g., ones from race to crime). The alleged disregard for individual autonomy in profiling may partly stem from a confusion between epistemology and metaphysics. When statistical discriminators enact on correlational evidence, this is taken to express a harmful or demeaning metaphysical claim about the discriminatees:

> For a profile to be useful, the possession of a particular trait must only raise the *epistemic* probability that a person is a criminal. But the distinction between a trait's making one (epistemically) more likely to be a drug courier, and its actually *making* one more likely to be a drug courier, is predictably lost on many people. (2015, 210.)

What Eidelson implies is that statistical generalization may be (falsely) understood as disrespectful in the basic sense, as if statistical evidence would to express a categorical or metaphysical claim about people of color and criminal propensity, thereby undermining their autonomy. In chapter 4, I suggest that the fact that profiling bears a specific social meaning may be, nevertheless, relevant to the moral evaluation of profiling – and is so even if we concede that correlational "evidence is only relevant as evidence", in Enoch & Fisher's terms (2015, 569.). The experience of being scrutinized on the basis of a profile may be qualitatively different for individuals of different ethnic background in a society characterized by tensions related to race, for example (cf. Eidelson 2015, 196–197).

3.2. Unfairness

A second principled claim is that wrongful instances of algorithmic discrimination involve unfairness understood as unequal treatment or unjustified partiality. As noted in chapter 2.3.2., fairness has indeed become a key question in AI ethics and fair ML. The discussion on fair ML ranges from issues with predictive accuracy to the permissibility of using sensitive characteristics as a basis for differential treatment. As such, it echoes claims raised also against racial profiling[104]. Allegedly, it is wrong to differentiate treatment based on information about people's traits that are not (i) chosen by

[104] These objections are considered, e.g., by Eidelson (2015, ch.6.2.1.), Lippert-Rasmussen (2007, 397–399) and Thomsen (2013).

them or (ii) cannot be altered. Following this line of thought, when ADSs compute sensitive information explicitly or when their predictive proxies roughly align with sensitive group-membership, they will wrongfully discriminate. Next, I consider these claims and argue that they do not provide an exhaustive explanation of the underlying wrongness of algorithmic discrimination.

3.2.1. *Fairness, Redistributive Justice and Conflicting Ethical Principles*

Recall that evaluations of overall model accuracy tell one little about fairness because the validation data contains the same biases as the training data on which an algorithm is trained (see ch. 2.3.2.). Fairness considerations have their limits as well, however. This is exemplified by the debate on the alleged racial bias in the outcomes of the risk assessment tool COMPAS. An independent non-profit journalistic organization, ProPublica, famously released a study (Angwin et al. 2016) according to which the rate of false positives COMPAS produced was significantly higher for black defendants (44,9 %) than for white defendants (23,5 %). False negative rates were also disproportionate: 47,7 % for white defendants and 28,0 % for black defendants. (Dressel & Farid 2018, 1.) COMPAS seemed "to favor white defendants over black defendants by underpredicting recidivism for white and overpredicting recidivism for black defendants" (Ibid.). Critics have since insisted that COMPAS' predictive accuracy ought to be equal with the respect to both groups. According to this fairness metric, predictive equality, false positive rates should be equal across racial groups (Corbett-Davies et al. 2017, 798). However, the company behind COMPAS, Equivant (former Northpointe), released a refutation of ProPublica's study, claiming that COMPAS' predictions were equally accurate with respect to both black and white defendants. Insofar as "the correct classification statistics are used, the data do not substantiate the ProPublica claim of racial bias towards blacks", they claimed (Dieterich et al. 2016, 1). As it turns out, by correct methods Equivant meant employing a different fairness metric. Equivant demonstrated that once defendants' statistically evinced likelihoods of reoffending were taken into account, the accuracy of the predictions with respect to individuals in both groups were in fact roughly equal (Dieterich et al. 2016). COMPAS was demonstrated to be similarly calibrated across groups[105] – i.e., irrespective of a defendant's race, the probability that a given defendant would go on to reoffend was the same for any given risk estimate it produced (Dallas 2018; Dressel & Farid 2018, 1).

[105] Well-calibration is an extension of *calibration*, which is satisfied when "for any probability score [*s*] subjects in both protected and unprotected groups have equal probability to truly belong to the positive class" (Verma & Rubin 2018, 5) – e.g., receiving a positive loan decision when they have an actual positive value, e.g., are actually creditworthy. Well-calibration, then, requires that this probability is also equal to *s*; the model's predicted probability score *s* in relation to some group should be approximately equal to the number of positive actual values with respect to that group. (Verma & Rubin 2018, 5.)

The ProPublica v. Northpointe debate exemplifies an integral finding in the literature on fair ML: some fairness metrics and definitions, with their respective benefits and faults, are mathematically incompatible.[106] Kleinberg and colleagues (2016) demonstrate that a classifier algorithm, such as COMPAS, cannot simultaneously satisfy the metrics of well-calibration, balance for positive class, and balance for negative class, except in highly constrained cases. Notably, if a model satisfies the metric of balance for negative class, it will necessarily satisfy predictive equality as they are mathematically equivalent (Verma & Rubin 2018, 4). As it turns out, the measures of fairness employed by Equivant and its critics are incompatible. In COMPAS' case, this incompatibility becomes pertinent due to the baseline differences in the data: people of color recidivated at higher rates in comparison to white people (cf. Dressel & Farid 2018). Calibrating the system accordingly results in an increased likelihood of black defendants being predicted as high-risk for recidivism, as traits correlating with one's race were found to be predictors for reoffending.

The general dispute in the literature over what fairness metrics should be employed may be partly explained by tensions between the implicit assumptions about equality, fairness, and non-discrimination that underlie those definitions to begin with. Indeed, Binns argues that the term "'fairness' as used in the fair machine learning community is best understood as a placeholder term for a variety of normative egalitarian considerations" (2017, 2–3). Depending on the context of use of an ADS, he says, some approaches to fairness may more apt than others as they will correct different "kinds" of inequalities:

> Equality of opportunity may be an appropriate metric to apply to models that will be deployed in the sphere of 'economic justice'; e.g. selection of candidates for job interviews or calculation of insurance. But in contexts which fall under the sphere of civil justice, we may want to impose more blunt metrics like equality of outcome (or 'parity of impact'). (Binns 2017, 7.)

According to Binns, then, depending on the context and the "equality of what"[107] one is interested in when using an ADS, different formulations of fairness may prove more suitable than others. In this sense, fairness metrics introduced in the fair ML literature are not only about (un)equal treatment, but also involve considerations regarding redistributive justice.

This in mind, I would argue that, as COMPAS satisfied the metric of well-calibration, Equivant was correct in saying that it treated everyone *equally*. This is true despite both the fact that

[106] For an in-depth examination on the subject, see Friedler et al. (2016).
[107] The "equality of what" discussion is an open debate in political philosophy, which concerns the question of how equality should be conceptualized. In other words, it revolves around the question as to what are the "dimensions within which the striving for equality is morally relevant". (Gosepath 2011, S.3; see also Sen 1979)

its accuracy was poor overall[108] and the fact that well-calibrated models will not work well on an individual level[109]. Regardless of these general flaws, the same thresholds for risk-scores were indeed employed for everyone. (It is a category mistake to assume that unjust or otherwise poor policies could not treat individuals equally – one could, for example, conduct medical diagnosis by flipping a coin.) Due to differences in (discovered/evinced) reoffence rates, however, black defendants were more likely to belong to the high-risk group as a result of employing the particular threshold. Accordingly, I suggest that ProPublica's argument from unfairness should, in fact, be understood as a disguised demand for *redistributive justice*. Claiming that COMPAS is unfair, ProPublica were in effect calling for what Chander (2016) terms "algorithmic affirmative action". Chander says that insofar as "we believe that the real-world facts, on which algorithms are trained and operate, are deeply suffused with invidious discrimination", the solution is to "design our algorithms for a world permeated with the legacy of discriminations past and the reality of discriminations present" (2016, 1025). Redistributive justice (through predictive equality) would in COMPAS' and other formally similar cases require employing *different* risk thresholds for different groups – i.e., unequal treatment.

Thus, when ProPublica and other critics demand equal predictive accuracy in COMPAS' case, they are in the business of correcting past wrongs as opposed to pointing out unequal treatment or wrongful discrimination as such. However, when a group is comparatively disadvantaged in virtue of higher (or lower) evinced baseline propensity for some attribute pertaining to that group, this is not necessarily "unfair". Taken as a categorical claim, it would mean that it would be *unfair* (or rather, unjust) to focus, say, preventive mental health services to those that are most at risk (i.e., to calibrate such a policy according to evinced risk factors). Nor may it be the most desirable option in all cases to employ different risk-thresholds for different groups. Demanding equal false positive and false negative rates in COMPAS' case would conflict with the aim of achieving maximum security (i.e., minimizing the set of violent crimes) because this would increase the overall false negative rate of the model (Corbett-Davies et al. 2017). In this sense, even if demanding redistributive justice might be the *just* thing to do in some cases, COMPAS (or other ADSs alike) are not necessarily unfair even when their predictions have a disparate impact on some group. We also cannot

[108] Notably, COMPAS' predictions which were – to the best of our knowledge – based on 137 variables were not in fact significantly more accurate than those of laymen with only 7 predictive variables at their disposal. Comparing COMPAS' predictions against those of human participants across 20 subsets of data, Dressel & Farid found that "the average of the 20 median participant accuracies of 62,8 % […] is, just barely, lower than the COMPAS accuracy of 65,2 %" (2018, 2). In other words, humans could achieve roughly equal accuracy as COMPAS by using only 7 variables. Thus, perhaps we could regard COMPAS as "unfair" or unjust for everyone in this sense. However, for the sake of an argument we can suppose that some similar system could accurately predict recidivism.

[109] Although well-calibrated models allow for confidence in predictions in the long run, they fail to account for nuances between individuals in a given group (e.g., between defendants in the high-risk group). In other words, a calibrated model assumes that any group *Y* that is *P* probable to reoffend will be homogenous. (Dallas 2018.)

straightforwardly dismiss that in similar cases, two morally significant goals may be in conflict – e.g., maximizing public safety and correcting wrongs both past and present.

Should we still insist that COMPAS be blind to race (or, rather, a rough proxy for it)? This seems problematic. Corbett-Davies and Goel note that because "women are typically less likely to commit a future violent crime than men with similar criminal histories", gender-blind models may "systematically overestimate a woman's recidivism risk and can in turn encourage unnecessarily harsh judicial decisions" in the context of criminal justice (2018, 2). Notably, here the differences in baseline reoffence rates differ not between racial groups but genders instead; if the set of predictors that function as a proxy for race are not taken into consideration, white defendants could suffer the same fate in colorblind models as women would in genderblind models.

It seems we have arrived at the fundamental question of whether (and when) we should remain blind to or be aware of people's sensitive traits in decision-making. This translates to a question of whether we should employ the similarity-based metric *fairness through unawareness* (see chapter 2.3.2), as opposed to *fairness through awareness*, which accepts that sensitive traits can be relevant for decision-making (cf. Verma & Rubin 2018, 6). Similarity-based measures are motivated by the fact that parity measures (e.g. statistical parity) may hide procedural unfairness when the training data suffers from sampling bias[110]. They build on the idea that similar individuals should be treated similarly, regardless of their legally protected attributes. This idea is echoed also by causal approaches to fairness. *Counterfactual fairness*, for example, requires that an individual's prediction or decision will remain the same in a possible world where her protected attributes are different – i.e., where that individual does not belong to a protected group (Kusner et al. 2017, 2).

A conundrum manifests itself. On the one hand, color- and gender-blindness seem appropriate as general rules for high-stakes decisions. Indeed, sensitive traits are commonly deemed as illegitimate grounds for differential treatment, and this idea underlies many of the fairness definitions considered above. On the other hand, in the recidivism case it seems we should appreciate the fact that women exhibit a lesser tendency towards recidivism (i.e., be aware of defendants' gender). As we saw, employing the definition of fairness through unawareness may even be harmful in cases where there are actual baseline differences between groups. But is there a principled argument against relying on information about sensitive traits in AD – irrespective of legal norms, that is?

[110] Verma & Rubin give the following example in the context of credit-scoring: "[S]uppose the same fraction of male and female applicants are assigned a positive score. Yet, male applicants in this [data] set are chosen at random, while female applicants are only those that have the most savings. Then, statistical parity will deem the classifier fair despite the discrepancy in how the applications are processed based on gender." (2018, 5.) As noted above, statistical parity requires that protected and unprotected groups are equally likely receive a positive decision.

3.2.2. *Objections from Responsibility and Immutability*

Referring to the case of COMPAS, Binns entertains the following idea:

> [O]ne of the potentially objectionable features of the COMPAS scoring system was not the use of 'race' as a variable (which it did not), but rather its use of variables *which are not the result of individuals' choices*, such as being part of a family, social circle, or neighbourhood with higher rates of crime. These may be objectionable in part because they correlate with 'race' in the U.S., but they are also objectionable more generally to the extent that they are not the result of personal choices. (Binns 2017, 7; italics added.)

According to this common view, discrimination is intrinsically wrongful if and when it relies on information about traits for which an individual is not responsible. Another view objects to discrimination on the basis of traits that one cannot alter. When decisions are based explicitly on information about individuals' "immutable traits", Zarsky notes, this may amount "to subordinating minority groups" and further "their seclusion" in different societal contexts. (2014, 1387) However, some have held that differential treatment on the basis of immutable traits will allegedly be wrongful irrespective of outcomes, i.e., intrinsically objectionable (cf. Eidelson 2015, ch. 6.2.1.).[111]

Eidelson's reply to claims according to which statistical discrimination involving sensitive traits is inherently wrongful is that it would seem to make no *intrinsic* moral difference what the predictor trait in question is. Talking about racial profiling, he says that the morality of imposing burdens in this regard "has nothing to do with the fact that people are not responsible for their *race*" (2015, 192). Indeed, two notions give us reason to doubt the objection from responsibility. First, responsibility for some trait does not seem to make an intrinsic difference for the morality of discrimination as some instances of such discrimination are regularly deemed innocuous (Mogensen 2019, 456). Policies that restrict individuals' opportunities on the basis of age are not always considered wrongful (e.g. right to vote), nor is it common to oppose height restrictions to rides at amusement parks, even though both age and height are traits that are not under individuals' control[112]. It would require further specification as to why discrimination on the basis of race would differ in this regard. Secondly, one may be treated wrongfully even if she *is* responsible for her trait(s). Innocent people who happen to look like stereotypical shoplifters, for example, are generally

[111] Distinguishing the arguments from responsibility and immutability is important as responsibility and choice may in fact not align in all cases Traits of which an individual is not responsible are not necessarily such that they cannot be altered (cf. Lippert-Rasmussen 2008, 397–398). An individual might be coerced by others to take on a mutable trait (e.g. one may be forced by others to wear religious attire that may trigger scrutiny). Conversely, an individual may be responsible of a trait that cannot be altered in the future (e.g. one may be responsible for missing a credit-card payment, and the fact that she has done so cannot be changed *after* the fact).

[112] Height, perhaps, can be surgically altered. For this argument it suffices, however, that this is generally not the case.

responsible for how they dress or conduct themselves. Assume now that such people are more often subjected to searches and disproportionately surveilled in stores in comparison to those not fitting the profile of a shoplifter. Should the fact that one chooses one's attire entail that she is treated fairly when subjected to disproportionate scrutiny? Eidelson thinks not[113]. Even if one were responsible for some trait, he argues, this does not entail that she is responsible *for the treatment* that she is subjected to on the basis of having that trait. (Eidelson 2015, ch. 6.2.1.) These are two different objects of responsibility. Thus, the notion of responsibility does not ground the distinction between morally neutral and objectionable discrimination.

The argument from responsibility seems especially problematic when one considers the opaque nature of profiling. Profiling and related practices (e.g. security screenings) are often conducted in secret because opacity concerning the factors that trigger scrutiny (i.e., differential treatment) is essential for those practices to work. When those practices are opaque, it will be irrelevant whether individuals have chosen to have some indicator trait, then, because they have not been granted a chance to *choose otherwise* in order to avoid the burden that comes with having that trait. In other words, knowledge of what exactly triggers some form of differential treatment seems to be relevant only to the extent that individuals can actually alter the trait in question. Transparency, Eidelson holds, does not make a moral difference insofar as one cannot choose whether to continue to bear the indicator trait or not. (Eidelson 2015, 189–190.)

The point about transparency seems to now shift the focus on whether one can alter her traits, the core of the second objection. Again, Eidelson thinks immutability is not the core of the moral issue. He argues that whether one can change some trait will matter only in relation to the potential costs and benefits that come with being treated differently on the basis of that trait:

> [B]urdening someone on the basis of a trait that she can change is not necessarily *at all* fairer than burdening her on the basis of other traits: it will be fairer only if the cost to her of abandoning or suspending the trait is less than the cost of the elevated chance of search. (Eidelson 2015, 191)

In other words, if one can choose whether to (continue to) have some trait (e.g. dress in religious attire), this will matter only if one knows that having that trait will involve a greater risk of having to bear some burden (e.g. being searched). Responsibility, Eidelson maintains, plays a role only in the following, contingent way: If one chooses not to change that trait, she will be (at least partly)

[113] Eidelson does concede that discrimination on the basis of chosen traits might be *less bad* to some extent, but this is not because they have chosen those traits. Rather, it will be so because the traits are likely to be such that they can be altered, shifting the focus on mutability. (Eidelson 2015, 190.)

responsible for the fact that the risk of being singled out has increased. Nonetheless, she will *not* be responsible for the burden that is imposed on her. (Ibid., 191.) Along these lines, if one wishes to dress like a stereotypical shoplifter (knowing that this will raise suspicion) one will be responsible for the increased risk of being searched, but not for the shopkeeper enacting on his suspicion.

The examination of fair ML so far points towards two notions: Firstly, all instances of wrongful algorithmic discrimination may not be unfair, although they might be wrong for some other reason. Secondly, in some contexts, morally just treatment may require treating different individuals similarly while, in others, we might have to engage in non-identical treatment of those same individuals in order to ensure equality. (Eidelson 2015 & forthcoming; see also Lippert-Rasmussen 2007, 396.) In other words, whether we should remain blind to individuals' sensitive traits or, conversely, appreciate the fact that they might have normative significance for our decision-making (i.e., be aware of them) seems to be sensitive to context. What matters for those considerations is whether there are baseline differences captured by data and, perhaps more importantly, what those differences *represent* (e.g. are they a result of past discrimination). In addition, we also need to consider what normative goals we are pursuing in making decisions – e.g., minimal threat and risk or redistributive justice (Binns 2007; Selbst et al. 2018).[114]

3.2.3. *Which Groups Matter? Two Forms of Gerrymandering*

One question regarding fairness persists: can we delimit the set of relevant groups with respect to whom AD processes should be fair *a priori*? Notably, Anna Lauren Hoffmann (2019) has criticized the current approaches to fairness due to their inadequate attention to intersectionality in discrimination[115]. Hoffmann argues that common approaches to fairness fail to recognize discrimination that may be compounded and multi-dimensional. Individuals may belong to multiple marginalized groups (which may or may not be legally protected). The "single-axis thinking" persisting in the current discussion on fairness overlooks how individuals are often "multiply-oppressed" (Ibid., 906). Mann and Matzner note similarly, that "algorithmic profiling that facilitates the inclusion of different sources and types of data is likely to contribute to increasing entanglements of protected identities, thus creating new categories and groups of people experience forms of

[114] This conclusion does not rule out the possibility that there *are* instances of unfair decision-making; it only denies that unfairness would ground the wrongness of *every* instance of wrongful algorithmic discrimination.

[115] According to intersectionalist thought (cf. Crenshaw 1991; Stoljar 2017) the disadvantage faced by a racialized woman, for instance, may be qualitatively different than that of a non-racialized woman. This may be the case even when both groups may be severely disadvantaged in comparison to some other, privileged group, such as (non-racialized) men. Stoljar (2017) has argued accordingly that, in evaluating whether an act is discriminatory, we ought not to focus only on socially salient groups but "normatively salient groups" as well. This is because some multiply disadvantaged groups might not actually be socially salient (e.g. LGBTIQA+ couples who live together).

discrimination", adding that "[i]ntersectional theory has shown that safeguards against discrimination wrongly assume that all forms of discrimination function similarly or independently" (2019, 5).

Indeed, statistical metrics for fairness, for example, tend to address differences between some pre-defined groups in a given population and in a binary manner; by comparing predictions between two, often legally protected groups (e.g. 'women' and 'men'). Kearns et al. (2019) argue that this can lead to so-called "fairness gerrymandering": The outcomes of algorithmic decisions may satisfy or approximate some measure of fairness with respect to these groups; e.g. an algorithm may be fair in that it is not significantly biased against women. However, cases where disadvantage results from a *conjunction* of two or more protected attributes remain unaddressed; e.g. 'black women' may systematically be treated worse than 'white women'. Intersectional discrimination produces a two-way problem for algorithmic fairness, then:

> There are exponentially many ways of carving up a population into subgroups, and we cannot necessarily identify a small number of these *a priori* as the only ones we need to be concerned about. At the same time, we cannot insist on any notion of statistical fairness for every subgroup of the population: […] any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it misclassified. (Kearns et al. 2017, 2.)

What we see is that, on the one hand, data miners cannot determine what groups (i.e., subpopulations) should be taken into consideration in fairness evaluations *a priori*. On the other hand, if we look at statistical measures, the model will be unfair for any group specified by a misprediction or misclassification. In this sense, it is a *normative* assumption that some groups (often legally protected ones) matter more than others when one is engaging in fairness considerations to begin with[116]. Any assumption in this regard will require further justification from the point of view of moral theory.

As it turns out, this point also echoes the Synthetic Groups Question. Fairness evaluations are often limited to binary considerations with respect to socially salient groups; thus, they overlook instances of intersectional discrimination involving multi-dimensional disadvantage. But they also overlook instances of what Mann and Matzner (2019) called emergent discrimination. Using ML models (ones extracted via unsupervised methods, in particular) may lead to discrimination against "synthetic" groups specified by mixtures of non-salient categories of identity – the only "trait" binding them together being the value of the target variable (or a negative decision). Fairness gerrymandering may thus not only hide discrimination against multiply oppressed groups, but also

[116] Notably, this normative assumption may in most cases follow anti-discrimination legislation, which specifies grounds for prohibited differential treatment. Nevertheless, as here the ethics of discrimination are under consideration – and not legality thereof – one cannot assume a given set of legitimate grounds *a priori*.

"synthetic" ones, which are not salient in the sense that they would commonly enter our ethical considerations when they are discriminated against.

One finds a point of emphasis here which relates to the notion that discrimination is actual-properties independent (see chapter 1.1.1): whether a prediction is *correct* or not does not tell us whether a group is *comparatively disadvantaged*. One could be comparatively disadvantaged on the basis of accurate predictions as well. For example, even if COMPAS' scores were accurate, it could still mean that black defendants would be denied parole more often than white defendants. And, given that many socio-economic inequities, such as poverty, could underlie such a tendency, this seems morally problematic. Conversely, one might enjoy comparative advantage on the basis of a misprediction. If white defendants are systematically and falsely predicted not to reoffend, they will enjoy comparative disadvantage in relation to black defendants. Wrongful discrimination, thus, seems to be irreducible to the problem of misprediction alone. The question is, rather, *why* one kind of discrimination would be morally justified while the other one is not. Fair ML cannot provide an answer to how can we avoid "objectionable-discrimination gerrymandering" – i.e., the SGQ.

3.3. Chapter Summary

In this chapter, I argued that the wrongness of algorithmic discrimination does not lie ultimately in the fact that individuals are treated on the basis of information about their respective (sensitive) group-status. This is because, while AD involves treating people on the basis of the generalizations, it will not intrinsically involve disrespect for individuality. First, I noted that objections against generalization often find fault in the use of inaccurate generalizations as opposed to generalizations as such. Second, the idea that statistical evidence does not justify discrimination – when causal evidence would do so – is not defensible: Reliance on statistical evidence may be a highly reliable practice that produces results. It is also a common one, even in contexts which involve high-stakes decisions, such as police work. Insofar as the causal connection argument goes, it is also implausible in and of itself, because it does not ground a distinction between individualized and other evidence. Thus, the causal connection argument does not exhaustively explain why profiling that involves reliance on information about sensitive traits *in particular* would be wrongful. Third, I argued that while ADSs might not capture each individual's normatively significant character and AD may not fully appreciate these individuals' agency due to a lack of deliberative flexibility, this may be due to issues with data accessibility and thus does not necessarily constitute objectionable discrimination. I also noted that even if AD were to involve disrespect for individuality, it may not necessarily involve *comparative* disrespect for individuality.

I argued that the notion of algorithmic fairness (or discourse thereof) fails to address some underlying moral issue with algorithmic discrimination. Fairness considerations always involve normative assumptions regarding both the "currency" of equality and the set of groups that are deemed relevant stakeholders. In this light, I noted that objections from unfairness are sometimes confused with demands for redistributive justice. These demands may be justified but nevertheless end up confusing wrongful algorithmic discrimination with unfairness (understood as unequal treatment). Furthermore, I noted that the problem with algorithmic discrimination is not exhaustively explained by reference to responsibility for traits or mutability thereof, although the idea that these would explain the wrongness of discrimination is common. I concluded that whether reliance on sensitive traits is objectionable or, perhaps even desirable, is dependent on context and purposes of a given decision-making procedure.

I have suggested that the problem in relation to autonomy considerations seems to relate to *what* a given statistical generalization expresses in relation to social context. I have also hinted towards the notion that one should distinguish objections from unfairness from ones "that simply indict profiling as especially harmful" (Eidelson 2015, 196). The harm done by algorithmic discrimination alongside the social meanings it may express, I suggest, provide a starting point for countering the Nothing Personal Argument and answering the Synthetic Groups Question. Before I attempt to flesh this out, two clarifications ought to be made: Firstly, my examination does not entail that causal evidence would not generally be more preferable or that predictive accuracy would not matter. I have only suggested that reasonably accurate statistical evidence is a justified basis for differential treatment, even when it concerns people's sensitive traits. Second, my examination does not rule out the possibility that issues with objectification, inaccuracy and unfairness could accompany instances of wrongful algorithmic discrimination.

4. Contingently Wrongful Algorithmic Discrimination

I have argued that *unalloyed* algorithmic discrimination will not be necessarily wrongful – at least on grounds of the arguments that have been considered throughout this study – even when it has a disparate impact on protected or vulnerable groups. This gives us reason to suspect that the wrongs involved in statistical discrimination (and algorithmic discrimination, respectively) are contingent in nature. That is, such practices will be wrong (when they are) in light of their consequences. I will next consider Eidelson's Broad Harms Argument as a way of explaining the wrongness of unalloyed algorithmic discrimination. I start by considering Eidelson's argument in the context of racial profiling. After this, I proceed to apply the argument into practices which involve AD, such as predictive policing, recidivism risk assessment and credit-scoring.

4.1. Expressive Harms and the Broad Harms Argument

Profiling and other practices involving statistical discrimination are often defended in terms of the benefits they may produce. Profiles offer better allocation of resources and enable targeting of limited demographics as opposed to conducting randomized searches, which may be more costly. Thus, depending on the context, it may prove a cost-effective option in serving some public good (e.g. security). Reliance on statistical evidence, although it will always involve a risk for individual mispredictions, will also tend to prove effective in the long run. Furthermore, acting on statistical evidence of crime patterns may not only prove effective but also be beneficial for members of each group (e.g. by reducing violence in communities) even if we acknowledge that the evidence reflects inequalities wrought by oppressive practices and discrimination both past and present. (Risse & Zeckhauser 2004, 140–143; Schauer 2006.)

This is not to say that profiling would not also involve significant harms. These may include both direct and indirect harms. Individuals subjected to scrutiny on the basis of their (sensitive) traits may feel humiliated or anxious, and perhaps experience feelings of resentment and shame. Call these *direct belief-dependent harms*. Profiling may result in *material harms* as well, such as inconvenience (e.g. delays at airports due to screenings) or economic costs, for example. Material harms fall on both the discriminator (e.g. costs of implementing profiling practices on the field) and those discriminated against (e.g. missing a flight due to scrutiny). In addition, one might distinguish what may be called *interpersonal harms*. These are belief-dependent harms that are produced when the subject is degraded or demeaned or stigmatized in the eyes of others due to being singled out in

some way – e.g., when one's bags are searched at an airport and this is witnessed by others. (Eidelson 198–202.)

Few clarificatory notes should be made at this point. First, the set of people who believe they are being scrutinized on the basis of some property and the set of people of who are *actually* scrutinized on the basis of that property ought to be distinguished; these sets of people do not always fully align. One may either falsely believe that he is being profiled P-wise while he is not. Conversely, one could be profiled P-wise while he may not know it. (Eidelson 2015, 198–202.) In the context of racial profiling, for example, this means that one "should be wary of understating the harms of adopting a racial profiling policy by assuming that they are suffered only by people whose targeting is *actually* influenced by race" (Ibid., 200). Furthermore, if a trait is significant for an individual (e.g. it contributes greatly to her identity), the belief-dependent harms may supposedly be more severe when she is scrutinized on that basis. In other words, one might be more opposed to being singled out on the basis of one trait (e.g. race) than another (e.g. height). (Ibid., 198–202.)

Considering how one should evaluate the morality of racial profiling, Risse & Zeckhauser (2004) famously introduce the notion of *expressive harms*. Expressive harm, they say, "occurs primarily because of harm attached to *other* practices or events" and it is produced when a policy or practice functions as "a *reminder*" of past and present injustice and "other painful events or practices" (Ibid., 146–147). It may also occur when that policy or practice bears a symbolic meaning "of structural disadvantage or maltreatment" (Ibid., 147). Risse & Zeckhauser see expressive harms as the prominent reason why people often object to statistically discriminatory practices, such as racial profiling. They motivate this claim with the following thought experiment:

> [I]magine the closest possible world to U.S. society except that there is no racism. Race-caused disparities in economic or educational attainment do not exist, and practices such as race-related police abuse are unknown. In such a society, we conjecture, using race for investigative purposes would not be considered offensive and would not trigger resentment, hurt, or loss of trust in law-enforcement. (Risse & Zeckhauser 2004, 146.)

Expressive harms, in this sense, are contingent upon the social and historical context in which different practices take place. When one is singled out on the basis of her sensitive traits, this may echo other injustice that a person or others like her have faced in the past. Belief-dependent and intrapersonal harms may most often be expressive in this sense, although this is not necessary[117]. The

[117] Supposedly, one can feel anxious or demeaned when subjected to scrutiny on other grounds as well (e.g. when one gains unwanted attention). One could also be degraded in the eyes of others for reasons than what the scrutiny expresses in relation to some social and historical context.

notion of expressive harms will also perhaps explain why profiling is not objected to in cases where the discriminatees do *not* belong to a group that has suffered from systematic oppression. When young males are subjected to higher insurance premiums due to evinced propensity for traffic accidents, for example, this supposedly does not connote a history of injustice suffered by that group.

What does the notion of expressive harms mean for evaluating the morality of statistical discrimination? Firstly, proposing a utilitarian account of the morality of profiling, Risse and Zeckhauser (2004) maintain that one ought to weigh the *incremental* harms of profiling against its benefits. They emphasize that the relevant harms under consideration in each case should not include those related to what I called background injustice – i.e., ones that precede the act of profiling itself, such as historical oppression of minorities. Only the incremental harms – e.g. ones that result from solely from racial profiling – ought to be accounted for in the moral evaluation of a given practice of statistical discrimination. Secondly, the notion of expressive harms calls for a context-sensitive evaluation of the morality of statistical discrimination. The magnitude of expressive harms produced in engaging in profiling, for example, will depend on the states-of-affairs that obtain in a given society. Thus, "in any given instance" of profiling (or other practice of statistical discrimination), "one will normally have to undertake a textured analysis of the actual psychological effects and social meaning of *that* practice" (Eidelson 2015, 201).

The central argument Risse & Zeckhauser (2004) make, then, is the following: If the harms associated with profiling are expressive – that is, harms that are not caused by profiling *per se* but, rather, "triggered" by it – we will most often have no sufficient objection to profiling. In other words, Risse & Zeckhauser claim that people often confuse expressive harms that are caused by profiling in relation to background injustice to the harms of background injustice itself. In essence, they propose "a causal thesis, which claims that the harms of profiling are generally caused by factors other than racial profiling" (Lever 2007, 21) If this is so, Risse and Zeckhauser argue, profiling will actually be more beneficial than it is harmful (although this will depend on context) insofar as we concede that the harms of background injustice itself should be counted as harms of profiling *per se*[118]. (Risse & Zeckhauser 2004; Eidelson ch. 6.3.). Notably, Eidelson takes what he calls Risse & Zeckhauser's "Narrow Harms Argument" to severely undermine both the magnitude and scope of expressive harms[119]: While he concedes that one should not count so-called background injustice in

[118] Even if we were to stop practicing (racial) profiling – supposing that the underlying unjust dynamics in society were the same – they say, we would not significantly decrease the amount of injustice prevalent in society (Risse & Zeckhauser 2004, 149).

[119] This interpretation is evidently supported by Risse & Zeckhauser's conclusion, according to which "profiling is an expressive harm in a race-conscious society, but the incremental injury profiling imposes beyond such wrongs as police abuse and racial discrimination is small" (2004, 169). They also claim that the harms resulting from profiling in itself are "comparatively modest" in light of the benefits it will bring in preventing and reducing crime, for instance (Ibid., 149).

itself as moral cost of profiling, he maintains that "there is every reason to think that […] the elimination of racial profiling *would* constitute a significant improvement in the well-being of the people affected" (2015, 205). That is, even if the most salient harms of profiling were in fact expressive, this will not mean that they are not significant harms. Furthermore, it seems they do not concern only those discriminated against; they stem from, and spread across, wider social relations between individuals and groups. This is why Eidelson proposes an extension of the "Narrow Harms" view; the Broad Harms Argument.

According to the Broad Harms Argument, we should not consider only direct harms and individual, belief-dependent (expressive) harms when evaluating the morality of statistically discriminatory practices. Eidelson proposes that we extend the scope of relevant considerations from direct harms of statistical discrimination to broad, social harms to which those practices contribute by (re)producing and sustaining immoral behavior. He elaborates that profiling and other types of statistical discrimination – depending on the context – may produce (i) material harms, (ii) belief-dependent harms and (iii) interpersonal harms, as considered above. In addition, they may result in what I call (iv) broad social and reactive harms. They refer to "the more diffused consequences that result from institutionalizing and legitimating certain attitudes and habits of thought" (Eidelson 2015, 206). Racial profiling, for example, "may encourage the familiar and hurtful biases that attribute differences in rates of crime to deep-seated differences internal to the people involved, rather than to their social contexts" (Eidelson 2015, 210). It insinuates that ethnic minorities would be "perpetrators, rather than […] victims of crime" (Lever 2007, 23) and furthers segregation in society. In the context of policing, these diffused consequences may consist in racialized individuals feeling the need to avoid confrontations with the police; them being perceived as dangerous or suspicious by the community; and fostering distrust between groups of people in that community. Importantly, broad harms may be suffered by individuals or groups that are *not* disproportionately subjected to scrutiny and the notion of reactive harms accounts for this as well. Tension between groups may affect members of both minority and majority groups by leading to reactive and violent incidents across ethnic boundaries, for example. (Ibid., 205, 208–215.)

How does the Broad Harms Argument relate to the notion of disrespect, then? Expressive harms can be thought of as connecting to the notion of conventional disrespect. Singling out an individual on the basis of statistical evidence about a group she belongs to manifests disrespect prominently due to the social and historical context that grounds the relevant respect-conventions. Expressive harms are thus a result of violations of respect-conventions. This notion seems to vindicate the tension pertaining to the alleged disrespect that is manifested when statistical evidence is used to differentiate treatment between people: When an individual is singled out on the basis of statistical

evidence, the act may *express* that subjects are not autonomous individuals. The statistical evidence in itself, however, does not make any individual metaphysically more or less likely to be or do anything the evidence indicates, nor is it necessarily unfair or epistemically unwarranted for decision-makers to rely on such evidence. However, as Eidelson notes, "inferences from race to crime are so redolent of the profound contempt that has long underlain them, and thus are so likely to cause deep hurt, that they should not be used anyway; that is, their social meaning rules them out-of-bounds even in circumstances in which they may be epistemically unimpeachable and not […] disrespectful in the basic sense" (forthcoming, 50–51). Furthermore, an essential part of the argument is that racial profiling does not only express a conventionally disrespectful message and thus produce harm, but it also contributes to the production and reproduction of wider patterns of *basic* disrespect in society. In his words, "profiling is *contingently* bad; but among the ways it is contingently bad is that it will predictably induce people to act in ways that are *intrinsically* bad" (Eidelson 2015, 214). Thus, both the direct harms and broad social consequences of it should be weighed against its benefits. (Ibid., 208–214)

4.2. The Broad Harms of Algorithmic Discrimination

The Broad Harms Argument seems a plausible account of why one would oppose to practices that involve unalloyed statistical discrimination on the basis of socially salient traits, such as gender. When generalizations concern systematically disadvantaged groups, the social meaning expressed by discrimination is rendered conventionally disrespectful by virtue of it connoting injustice and subordination of those groups. It not only produces substantial harms but also symbolizes, sustains and amplifies attitudes and social practices that manifest basic disrespect. Yet the account allows for the possibility that statistical discrimination – an accurate and cost-efficient method for allocating resources and distributing goods in many cases – can be morally permissible in many contexts. In other words, the Broad Harms Argument provides support for a plausible account of the morality of statistical discrimination by identifying salient instances of objectionable discrimination without shoehorning formally similar, reasonable and seemingly legitimate practices into that category.

I will next consider the application of the Broad Harms Argument in the context of AD by using predictive policing as a primary example. I suggest that, in light of this argument, predictive policing and other "morally biased" uses of AD that have a disparate impact on some demographic group, may constitute morally objectionable algorithmic discrimination. Again, for the sake of my argument, I limit my examination to *unalloyed* algorithmic discrimination.

4.2.1. *Benefits of Algorithmic Discrimination (in an Unfair Society)*

To dig into the moral evaluation of predictive policing, consider first the material benefits of AD. Prominently, a major reason why organizations, including police departments, adopt data-driven approaches to decision-making is cost-efficiency. Data mining can offer them actionable insight to guide various tasks and operations: predictive policing algorithms can generate lists of potential perpetrators (or "heat lists"), map out possible areas where crime may occur, and create profiles of suspects. Automating these tasks may be more effective and cheaper (at least in the long run) as they streamline organizational processes and allocate resources more efficiently than unautomated and non-data-driven approaches. (Ferguson 2016; Selbst 2017.) Moreover, implementing ADSs is becoming more cost-effective as the necessary tools, such as sensors and computing power, are becoming increasingly cheaper. (Kelleher & Tierney 2018; Alpaydin 2016, 166–167.)

Consider now the belief-dependent, intrapersonal and communal benefits of predictive policing. When sufficiently accurate, the use of predictive policing algorithms may be effective in detecting and preventing costly risks and events that might otherwise prove costly. For example, even though predictive policing has been found to have a disparate impact on marginalized communities in the U.S., it has been reported to have significantly reduced crime in some areas within the time span of one year. In Santa Cruz, the use of PredPol resulted in an 11 % decrease in burglaries and a 27 % drop in robberies. In Chicago predictive policing algorithms predicted over 70 % of gun violence incidents. (Selbst 2017, 114–115.) While studies regarding the effects of predictive policing have been inconclusive in many cases, preliminary results suggest that adoption of the technology has led to a decrease in other types of crimes as well, including car thefts and property crimes (Ferguson 2016, 1130). If these outcomes are valid, AD solutions in crime prevention and security may arguably be beneficial for both the organization (e.g. in terms of material costs), but also for the communities within which they operate in that they ensure safety and promote a sense of security. Also, assuming that crime does not only occur across ethnic boundaries, predictive policing may also prevent violence within the communities and neighborhoods in which the preventive measures and interventions are more heavily targeted.

Thus, there is an argument to be made that predictive policing and other forms of AD may in fact benefit communities despite their possible disparate impact – including members of the group that the ADS is systematically biased against in its outcomes. However, these benefits may be partly explained by the fact that the predictive algorithms in such cases operate in a society characterized by inequality and unfairness. Plausibly, the fact that crime can be predicted through algorithmic means in, say, low-income areas in the first place, may be due the fact that inhabitants of these areas suffer from poverty and enjoy unequal opportunities in life or in that they have been

historically targeted by the police (cf. Benjamin 2019, 80–84.). This is not to say that algorithm-informed crime prevention would not be a reasonable goal; only that accurate predictive policing is possible partially due to the historical and social conditions that contribute to the occurrence of crime in the first place.

4.2.2. *Material Harms*

Let us next consider the material harms of algorithmic discrimination. While I am skeptical about finding a reasonable argument against algorithmic discrimination from the point of view of cost-effectiveness, it is still noteworthy that adoption of a data-driven approach comes with some material cost for a given organization. The development of algorithmic systems requires not only singular investments on resources data science projects, but also long-term investments in the regular evaluation, updating and refining of the systems. The costs of implementing AD are, of course, dependent on context, the complexity of the technological system and other factors. In any case, it is noteworthy that these costs are continuous due to the need for continuous evaluation and refinement (cf. Kelleher & Tierney 2018). AD will also involve costs for both the organization and the subjects of algorithmic decisions, which are relative to the accuracy of the model. Assuming that even sufficiently accurate systems will produce false negatives and false positives, these will be more or less costly and harmful depending on context[120]. Negative feedback-loops and negative spirals also increase the potential material damage for subjects of algorithmic decisions: (false) negative decisions, such as loan applications denied by an algorithm, may both feed into future AD processes where they function as a self-validating prediction of future negative outcomes, further restricting marginalized groups' access to goods. They may also send individuals down paths where data flows through networked systems that govern different aspects of their lives, effectively deepening their financial (and social) insecurity.[121]

4.2.3. *Direct Belief-Dependent Harms*

Now, consider the belief-dependent harms of AD, which may arise when individuals know or think they have been subjected to profiling. First, a point regarding the role of sensitive versus proxy classification in AD: The account offered here suggests that the demeaning effect of algorithmic discrimination is also connected to *beliefs* regarding bias in algorithms, and not merely whether or to

[120] Supposedly, false negatives come with potentially significant costs in domains such as crime prevention and public security where unidentified threats (e.g. terrorist attacks) may result in multiple casualties and property damage. In other more isolated contexts, false negatives may prove less costly for the general public (e.g. when an individual is misclassified as high-risk of defaulting as a result of a credit-scoring process).

[121] One could argue that the environmental costs of ML and related technology should be considered as material costs of AD practices as well. For an examination of the environmental costs of AI technology, see Dobbe & Whittaker (2019).

what extent a given algorithm actually computes sensitive information *per se*. Notably, Zarsky argues that the use of explicit information about individuals' protected group-membership "in a scoring process has a degrading effect, regardless of the negative impact of the detrimental treatment and to the extent its usage is publicly known" (2014, 1386–1387). I would say it is correct that, if an individual knows (or merely believes) that she is being treated differently on the basis of information about her race, this may be taken by the individual to express a demeaning message due to its racist social meaning. As I suggest below in chapter 4.3, Zarsky is also right to note that the transparency of sensitive classification is a pre-condition for the production of belief-dependent harms – the degrading effect may not follow when sensitive classification is opaque. However, it seems that belief-dependent (expressive) harms may be produced even when there is a proxy for a sensitive trait, such as race – i.e., when an individual's presumed race is redundantly encoded in "neutral" data. This is because one may not know whether an algorithmic model calculates a sensitive aspect of one's identity *per se* as opposed to a mere aggregate proxy, consisting of many "neutral" variables, such as ZIP codes. However, if the concrete outcomes (e.g. prediction rates and actions based thereupon) manifest or are alleged to manifest moral bias, an individual subjected to an algorithmic decision might conclude that her sensitive group-membership has illegitimately affected the decision.

Deviating slightly from Zarsky's statement, I also suggest that the possible degrading effect is closely connected to the context where AD is employed and the trait in question. Plausibly, most people would not oppose to using an individual's gender as an explicit factor in algorithmic medical diagnosis. It is a predictor of a range of health issues and thereby it will help in determining the kind of help an individual may need. But say a gender variable were used in an algorithmic assessment of a candidate's suitability for a job. In many cases this would raise concerns of gender discrimination. In employment contexts where competence and suitability for a job are (at least ideally) deemed more relevant than gender, sexist differential treatment may denigrate those who are disadvantaged as a result. Essentially, this point serves to emphasize the context-sensitivity of fair ML as discussed in chapter 3.2.: In algorithmic medical diagnosis, it could be that we prefer an ADS to be "aware" of our sensitive group-membership (i.e., dissatisfy the definition 'fairness through unawareness'). But in employment contexts, we could want the opposite; for our gender to remain unconsidered.[122] Thus, it seems that algorithmic fairness (in a broad sense) is connected to prevalent

[122] Naturally, the actual distributions in the data matter as well, as Corbett-Davies and Goel's example showed in chapter 3.2.1. The baseline prevalence of some medical condition relative to gender is arguably valuable information in medical diagnosis. Yet one might oppose to using data underrepresenting woman candidates in hiring decisions, when that data is used to train an algorithm for recruiting purposes. This might seem trivial a point, but it is less so if one appreciates the fact that in both cases gender may be predictive of some target trait (diagnosis or hiring decision) and there may be a formally similar distribution in the data with respect to the target variables.

respect-conventions, the context of application and the relevant 'currency' of equality in each context, as Binns (2017) argued.

Predictive policing, one could argue, can be considered a context prone to producing expressive harms when it has a disparate impact on people of color because of the meaning such a practice has in relation to race. In addition to producing material and psychological harms to those dealing with the police as a result, predictive policing may corrode individuals' sense of security and violate their dignity. The Stop LAPD Spying Coalition has interviewed community members of a historically over-policed neighborhood, Crenshaw in Los Angeles, regarding their views on the matter of profiling and predictive policing. The qualitative experiences of these community members are represented in a survey study conducted by the Coalition (N=289), which suggests that a majority of the respondents perceive the LAPD as untrustworthy and more than three quarters of the people surveyed suspected that they have been profiled on the basis of sensitive traits.[123] (Stop LAPD Spying Coalition 2013, 15–19.) The psychological harm suffered by members of the community are also reflected in the stories they told to the Coalition interviewers. One interviewee, "Ana E", states that the implementation of pre-emptive policing practices has effectively rendered her community unsafe:

> …they [the police] have resorted to simply being in our presence at all times of the day, everyday, in hopes to harass us psychologically and instill fear into us as we try to move on with our normal lives. They have resorted to the use of terror tactics to harm our family in any way possible that still allows secrecy from the general public. They have robbed us from our freedom, by taking our right to go anywhere without them stalking us. (Stop LAPD Spying Coalition 2013, 21.)

The experiences of community members that have been brought to light support the notion that, for members of historically over-policed communities and vulnerable groups, predictive policing may echo the sort of systemic injustice they face perhaps on a day-to-day basis, such as constant surveillance and police presence, pat-downs, abuse and police violence, and other forms of institutional racism.[124] Thus, implementation of an unalloyed yet racially biased practice of predictive

[123] When asked for one's "general overview of the LAPD", 180 out of 278 people said the LAPD "cannot be trusted at all", 93 said they "can be trusted sometimes" and 5 said they are very trustworthy. When asked "How often do you feel like you are profiled by law enforcement due to your race, religion, age and/or sexual orientation?", 82 out of 289 people said "very often", 125 said "sometimes" and 82 said "never". (Stop LAPD Spying Coalition 2013, 17, 19.) Notably, the survey is not peer-reviewed and lacks thorough explication of methodology. As such, its scientific validity and legitimacy may be called into question. Regardless, the aim of the study is to bring forth the qualitative experiences of members of marginalized groups and overpoliced communities and, as such, I would see it (and other studies alike) as valuable resources for gaining insight into how algorithmic discrimination may significantly affect marginalized communities.

[124] Furthermore, while these echoes and expressive harms can plausibly be exacerbated by knowledge that an algorithm computes race as an explicit factor, I doubt that the production of expressive harms would hinge on the specific way in which race figures into the explanation of the algorithm's output. In other words, racial classification may be *more*

policing can produce significant belief-dependent harms and undermine individuals' sense of security and dignity. Those individuals "might not know what algorithms are, but they know what it feels like to be watched"[125] (Benjamin 2019, 80).

4.2.4. *Interpersonal Harms (and the Feedback-Loops that Exacerbate Them)*

What about the interpersonal harms of AD? I argue that when algorithms and statistical models portray people of color systematically as high-risk individuals this can, in Eidelson's words, "be taken as validation for racist assumptions about minority groups, and there by nurture the very prejudices which most directly contribute to the oppressive relationships in question" (2015, 215). In other words, not only may algorithmic discrimination produce direct harms and express a conventionally disrespectful message, but it may also sustain and encourage attitudes and practices that are disrespectful in the basic sense[126].

Several theorists have considered the stigmatizing effects of algorithmic discrimination. Zarsky notes that "when a scoring system indicates that […] members of a protected group are less creditworthy", this expresses a message "which contributes to a historical prejudice that all members of this group are not trustworthy in many other contexts" (2014, 1399). Safiya Noble (2018) has offered a detailed account of how algorithmic systems sustain and exacerbate racism through algorithm-governed (mis)representation of people of color. She argues that by associating racialized individuals more likely with traits such as criminality and unprofessional conduct, or by portraying them in demeaning, sexually objectifying ways, these technologies maintain harmful racial stereotypes and privilege whiteness.[127] Similarly, Eubanks' (2018) case studies in sectors of public health and social welfare demonstrate how the implementation of AD systems not only traps individuals suffering from poverty into systems of constant surveillance and punishment, but also cements the social stigma associated with poverty. The "digital poorhouse" – a metaphor Eubanks uses to denote the use of novel technologies to surveil, police, and punish the poor – presents those suffering from access to basic goods and services and living below the poverty line as "undeserving" to the general public. These studies support the claim that statistical representation of different groups,

---

disrespectful in the conventional sense (because it is a clearer violation of respect-conventions) but whether racially biased predictive policing is disrespectful altogether is not dependent on *how* race figures into the logic behind an algorithm's decision, or so I would argue.

[125] The quote is from a Stop LAPD Spying Coalition workshop facilitator with whom Benjamin recalls chatting with.

[126] Notably, reproduction and amplification of disrespectful attitudes and conduct may follow irrespective of the *accuracy* of an ADS' predictions. Even if the predictions were based on sound indicators, those indicators may only be sound due to past systematic oppression and discrimination. Plausibly, racist and sexist attitudes may be amplified due to (cognitive) confirmation bias when subjects holding such attitudes are presented with information that supports them, regardless of the validity of that information.

[127] Specifically, Noble (2018) considers recommender systems, such as Google's search engine, and not ADSs as such. However, the general claim I make here applies in both cases.

and the predictions and decisions that subsequently follow, are not contextually isolated in their effects. Rather, they contribute to the stigmatization of marginalized groups, exacerbate segregation and alienation in communities and in society, as well as corrode social cohesion between groups.

It is also worth noting that the phenomenon of automation bias (see chapter 2.4.3.) is important with respect to interpersonal harms. "The scientific aura of the [algorithmic] scoring process will most likely further exacerbate the stigma-based concerns", as Zarsky notes (2014, 1401). Statistically evinced connections between protected group-membership and being "high-risk" or "undeserving of goods" function as deceptive, seemingly rational bases for differential treatment, although they may have been produced and reproduced through historical oppression. Automation bias may thus contribute to how algorithmic discrimination undermines conditions for respect between those enjoying privilege and those living in the margins. Bathing in an unhistorical and mathematics-based illusion of objectivity, algorithms create a deceptive justification for inequality and provide those with power and privilege "the ethical distance [needed] to make inhuman choices" (Eubanks 2018, 13).

Negative feedback-loops may also sustain and deepen the harms of algorithmic discrimination. As has been noted, algorithmic systems may trap data subjects into self-reinforcing and self-validating loops, where negatively affecting decisions are fed back into these systems in the form of input data. As a consequence, negative decisions lead to similar decisions in the future, deepening subjects' disadvantage and possibly restricting their future access to goods and services. This may happen in multiple domains due to the fact that data is not only fed back to a given system, but also often flows between different systems that govern different areas of societal life. Eubanks aptly describes the negative spiral dynamic which occurs when vulnerable groups are trapped into a networked technological system, reinforcing stigma attached to those groups:

> Marginalized groups face higher levels of data collection when they access public benefits, walk through highly policed neighborhoods, enter the health-care system, or cross national borders. The data acts to reinforce their marginality when it is used to target them for suspicion and extra scrutiny. Those groups seen as undeserving are singled out for punitive public policy and more intense surveillance, and the cycle begins again. It is a kind of collective red-flagging, a feedback loop of injustice. (Eubanks 2018, 6–7.)

This collective red-flagging may effectively corrode broader social conditions and intergroup relations when technological systems function to maintain the subordinate position of vulnerable groups, but also lead to reactive harms as a result of the social alienation of those groups.

However, as Zarsky notes, "the damages inflicted on individuals by the [negative spiral] dynamics are substantially different than those caused by the discriminatory practices" (2014, 1407). Even in the absence of systematic bias against marginalized groups, false predictions will almost inevitably harm some individuals (almost at random). The individuals sent down a negative spiral due to a *false* decision still have a claim that they are treated wrongly in such cases. (Arguably, one will have such a claim even when competent human decision-makers inevitably err.) Thus, the negative spiral dynamic is not a risk that concerns sensitive groups *exclusively* (Ibid., 1406–1407). However, due to the networked nature of digital technologies across multiple domains, disadvantage in one domain may transfer seamlessly to another, thereby deepening existing inequity which marginalized groups face in multiple areas of social and economic life. One should also note that the frequency of enacting a discriminatory practice does not entail unfairness of that practice. It may, however, impose *more harm* due to its frequency. (Eidelson 2015, 196–197.) Accordingly, while algorithmic discrimination is not necessarily "unfair" because some group is continuously and more often disadvantaged, it may indeed produce more harm because it may instantiate more frequently, partly due to the very nature of the technology in question. Thus, while feedback-loops and the negative spiral dynamic are not inherently related to discrimination and algorithmic bias (in neither the statistical nor the moral sense), they are artefacts of the connectivity between (socio-)technical systems which may effectively exacerbate the broad and reactive harms of algorithmic discrimination.

4.2.5. *Defeating the "Nothing Personal" Argument*

The Broad Harms Account seems to offer a way to defeat the "Nothing Personal" Argument and, therefore, explain why some instances of algorithmic discrimination picked out by the Structural Discrimination View may be wrongful. Relying on indicators that encode past and present injustice is not only possibly harmful due to material costs, but conventionally disrespectful due to the broad social and psychological harms it (re)produces. And this seems to be the case although one may not find fault in relying on statistical evidence in algorithmic profiling on grounds of disrespect for individuality, unfairness or the correlational nature of the evidence used to generate decisions. Insofar as the outcomes of AD practices are "morally biased" in that they trigger expressive harms in virtue of violating respect-conventions, this makes even well-intentioned use of reliable and accurate technology wrongful. In other words, while AD may be conducted with due care and respect, and by using reasonably accurate data and models, it may, in Eidelson's terms "undermine the conditions for effectively realizing a community of mutual respect" (2015, 214). Thus, unalloyed algorithmic discrimination may be contingently wrongful due to (perhaps, unpredictable or unforeseeable) harms resulting from it, or due to the intrinsically immoral conduct it contributes to.

My conclusion, then, is that not all wrongful algorithmic discrimination is traceable to (i) the fact that actors may act from discriminatory intent or manifest disrespect in the basic sense (Bad Actor View), nor to (ii) technology that relies on flawed statistical evidence or is inaccurate (Bad Technology View). Even well-intending actors with accurate and reliable technology may (re)produce systemic inequality and severe harms by basing practices and policies on models that capture statistical phenomena which are symptoms of past and present injustice. Furthermore, this account suggests that the fundamental wrong of algorithmic discrimination cannot explained by reference to whether sensitive information is explicitly operated on in AD, even though such conduct may in most cases violate respect-conventions or even possibly legal norms, such as anti-classification principles. Eidelson's account of autonomy suggests that sensitive classification is not intrinsically wrong, but this does not amount to a claim that such violations of social conventions were not wrongful altogether. Race-, gender- and other types of sensitive classification are prone to producing significant harms that may render even well-intended and otherwise reasonable or innocuous acts wrongful. Thus, in the context of algorithmic discrimination, this account entails that moral evaluation thereof should not only focus on the role of sensitive information, accuracy, or fairness in AD itself but also account for the benefits and harms that it may result in when embedded into different decision-making contexts that are characterized by distinct socio-technical contingencies, such as respect-conventions.

Lastly, it ought to be stressed that the Broad Harms Argument does not apply *only* to unalloyed instances of algorithmic discrimination, but tainted ones as well. In concrete practices and real-life applications, the moral issues with discrimination and those contingent to discriminatory practices may overlap to varying degrees. Instances of tainted algorithmic discrimination may be even more prone to do producing expressive harms as they also involve *other* moral violations. Algorithmic profiling may not only reproduce disadvantage, but it may be unfair or poorly conducted. Reflecting problems with unrepresentative or inaccurate data, members of a group may be systematically misclassified or have their actions mispredicted. In addition, algorithmic discrimination may amplify harmful and disrespectful practices that are nevertheless contingent to the use of algorithmic profiling in itself. Predictive policing, for example, may be accompanied by intrusions of privacy, violations of basic rights, illegal interrogation and detainment, harassment and threats (and not only related expressive harms), as the Stop LAPD Spying Coalition's study (2013) suggested. These immoral and illegal practices are enabled, sustained and amplified through the use of novel digital technologies, which may portray vulnerable individuals as dangerous or unworthy. As such, the moral evaluation of AD practices "in the wild" should account for the associated problems. Nevertheless, a more fine-grained conceptual dissection of these issues – one this study has tried to provide – serves to clarify the discussion at the level of moral theory. This is not only valuable in terms of theoretical clarity, but

also at the level of practice, because distinct moral violations may require distinct countermeasures, which I consider briefly below. First, however, I will suggest a possible answer to the SGQ.

4.2.6. *Answering the Synthetic Groups Question*

What makes emergent discrimination against synthetic, socially non-salient groups less problematic than those where an algorithm is "morally biased" and subsequently discriminates against, say, ethnic minorities or women? Eidelson's Broad Harms Argument may prove useful in explaining this peculiarity – i.e., answering the SGQ. Plausibly, algorithmic discrimination may not trigger expressive harms in absence of injustice associated with synthetic groups, which are specified by idiosyncratic (or) aggregate clusters of seemingly neutral traits, such as their choice of mobile phone brand or patterns in location data gathered from them. I would argue that these individuals may suffer narrow harms when scrutinized on the basis of such non-salient traits, similar to being scrutinized on the basis of one's shoplifter-*esque* attire in a convenient store may trigger feelings of shame and anxiety, for example. It is likely, however, that there is neither a cohesive nor salient identity binding members of synthetic groups together. Belonging to the synthetic group presumably contributes little to one's identity[128]; such groups are not ones "that the individual feels a strong affinity to, or that the public identifies as such" (Zarsky 2014, 1407). This is partly, perhaps, due to possible opacity regarding the cluster of traits that specify the group in question – i.e., individuals may not know *what* group they (are presumed to) belong to in the eyes of the algorithm. In this sense, expressive harms may not often follow to the same extent when compared to practices that involve algorithmic profiles where sensitive traits play a part, explicitly or via proxy. In absence of group stigma or histories of oppression attached to a synthetic group specified by an algorithm, the relevant psychological harms may not ensue (Ibid.). Furthermore, there may be no demeaning stereotypes pertaining to such groups, which would contribute to their stigmatization or alienation *as a group* in wider social contexts. Being targeted by algorithms supposedly will not contribute to hostile attitudes towards that group, even if a given individual could face such attitudes as a result of a negatively affecting decision. Synthetic groups are less susceptible to suffer from stigmatization, be systematically denied goods, or denied possibilities for engagement in many other areas of social life. Individuals of these groups may not suffer from existing vulnerabilities and inequity on multiple axes of social life, and they may enjoy wider social support and security than those in existing positions of vulnerability (Ibid.).

It seems, then, the Broad Harms Argument would provide a plausible answer to the SGQ. Indeed, given that many common practices, such as credit-scoring, medical diagnosis and risk

[128] Of course, people may be loyal to their mobile phone brand – e.g., Apple users may, perhaps, constitute a salient category of identity. I would argue, however, that the general notion I am defending here is plausible.

assessment in insurance, rely on statistical discrimination, this seems to be quite an intuitive notion. However, the fact that groups facing algorithmic discrimination are not so much specified by one or more socially salient traits, such as race and gender, and more so by a complex, aggregate clusters of traits discovered via data mining, may require reconsidering or refining current approaches to anti-discrimination law. The digital era may call for closer attention to intersectional discrimination that emerges through data mining and use of digital technologies. In fact, this has been noted in scholarship, exemplified by efforts to revise traditional lines of thinking with regard to legal protection from discrimination. Mann and Matzner, for example, note that "with respect to the abstracted nature of profiling and the drawing of inferences, it may not be possible to identify grounds for discrimination as per specific [legally] protected grounds, and that a broader and more diversified approach to anti-discrimination may be an avenue to explore" (2019, 3–4). In this sense, intersectional discrimination in the digital era presents a challenge for the identification of groups that are discriminated against as a result of data mining. Zarsky also notes that intersectional algorithmic discrimination, if prevalent, "might generate novel negative stereotypes and stigma attaching to social groups and personality traits" (2014, 1411). That is, algorithmic profiling could possibly lead to a development where intersecting, synthetic groups in fact become socially salient and, subsequently, become associated with possibly negative stereotypes. This is speculative, of course. Still, perhaps closer attention needs to be paid to the evolving dynamics of discrimination as a result of widespread implementation of novel data technologies.

4.3. The Dilemma of Secretive Algorithmic Profiling

The modified account of Eidelson's view I have presented in this study has proven at least a prominent candidate in explaining the wrongness of algorithmic discrimination. However, it faces a possible challenge or shortcoming related to transparency in practices of statistical discrimination. This problem is noted by Mogensen (2019, 459–460) as well as Eidelson himself (2015, ch.6.4.1.). Notably, it concerns Eidelson's view in general and not merely it's application into the context of AD. The problem is the following: If we assume that (i) algorithmic profiling may produce significant benefits (e.g., accurately prevent violent crime), (ii) these benefits are not outweighed by material harms resulting from profiling, and that (iii) the wrongness of unalloyed algorithmic discrimination is explained by belief-dependent, expressive harms, unalloyed algorithmic discrimination may be morally neutral or praiseworthy *given that it is conducted in secret*.

As expressive harms are belief-dependent, infliction thereof necessitates transparency regarding violations of respect-conventions. This entails that if (algorithmic) profiling is conducted

in secret, expressive harms may not arise, given that the relevant violations of respect-conventions do not come into light. On the one hand, individual belief-dependent harms may not arise if an individual does not know or believe she has been singled out as a result of algorithmic profiling. Similarly, interpersonal and communal harms are triggered only when such practices function as public symbols for injustice past and present, encouraging basically disrespectful behavior, which in turn necessitates that the biased outcomes of AD processes are publicly known at least to some extent. The dilemma is that if opacity in AD prevents the production of expressive harms, then it follows that otherwise seemingly harmful discriminatory practices may be rendered morally permissible only in virtue of being conducted in secret. From the point of view of consequentialist moral theory, this entails that the wrongness of (algorithmic) racial profiling would partly hinge on whether people (other than the discriminators themselves) know that such a practice is being conducted. Indeed, as Zarsky correctly notes, "the key to an extensive adverse impact on a specific segment of society is how visible and salient the process is in the eyes of those discriminated against, as well as to other segments of the public" (2014, 1400). Of course, all of this requires that the material harms of algorithmic profiling do not outweigh its benefits. I am inclined to say that in most unalloyed cases they will not, given that cost-effectiveness is a major reason why AD solutions are adopted in the first place (e.g., decreased expenses, efficient allocation of resources). The material harms are minimal when ADSs perform accurately, in particular, because of low false positive and false negative rates. Nevertheless, it seems controversial that secrecy and opacity would render morally biased predictive policing or other types of high-stakes decision-making morally neutral or perhaps even morally praiseworthy if it delivers results without the public's knowledge. Eidelson – while acknowledging what he calls "the most provocative consequence" of his argument (2015, 215) – is willing to bite the bullet, however. Even though racial profiling will be in most cases morally wrong according to his account, he concedes that it could be morally neutral (at least in theory) if it were conducted in secret. He emphasizes two points with regard to this admittedly striking conclusion.

Firstly, Eidelson claims that maintaining complete secrecy would "in nearly every actual case […] be exceedingly unrealistic" and that there is no "persuasive real-world argument to be made for secret racial profiling" (2015, 216). Thus, Eidelson is skeptical as to whether decision-makers could maintain such secrecy. I, in turn, am skeptical of Eidelson's response. The ubiquity and opacity of algorithmic systems may make secretive profiling more feasible, to say the least. Digital and AI-powered surveillance and the ubiquity of means for data collection provide organizations and decision-makers increasingly more sophisticated means for invisible governance, tracking and control. Law enforcement could run through databases in secret to find possible suspects and proceed to detain them without hassle, as the Immigration and Customs Enfocement (ICE) have effectively

done in the U.S (cf. Chan 2019). Indeed, lack of transparency currently encompasses AD both at the level of digital code and at the level of organizational decision-making, with public authorities' complex algorithmic tools and intended purposes thereof – including those of the police – remaining behind closed doors (cf. Malnick 2019). Thus, Eidelson's doubts regarding the feasibility of secretive policing, surveillance and opaque discriminatory practices may prove to be severe underestimations of the practical implications of this theoretical dilemma in the era of algorithmic governance. Moreover, Eidelson's response in this regard seems insufficient. The problem, as I understand it, concerns our intuitions about moral theory, and not only practice. Even if secretive profiling were impossible in practice, the theoretical implication may be problematic in itself. Thus, the dilemma of secretive algorithmic profiling should be taken to constitute at least a significant challenge for an adequate harm-based account of statistical discrimination.

Eidelson does give a second response as well. He notes that moral assessments and judgments regarding transparency can be distinguished from those regarding discrimination. Transparency may be desirable for many other reasons than (preventing) discrimination, he argues. It fosters trust between decision-makers and those they govern and thereby serves the principle of good governance. (Eidelson 2015, 216.) One could concede, then, that in the absence of actual (psychological) harm and disrespectful conduct (i.e., contingent moral issues) there really is no substantial moral violation to be found; at least one inherently related to discrimination. Indeed, if predictive policing were to *work* by actually preventing crime, increasing security and producing other benefits, all without reinforcing stigma associated to marginalized groups, should we not concede that it is morally neutral, at least? As a global review of AI ethics guidelines by Jobin et al. (2019) indicates, even in the context of AI, transparency is prominently seen as a "proethical condition" – it enables the realization of other ethical values and practices, such as principles of democracy, dialogue and participation, in addition to fostering trust and ensuring (legal) accountability. It is prominently also taken as a way to enable the prevention and mitigation bias in algorithmic systems and subsequent effects thereof, but it is reasonable to assume that transparency serves to prevent contingent issues related to algorithmic discrimination, in particular. These include, for example, data miners acting in malicious discriminatory intent and making poor, confirmation bias laden design choices.

One could concede that when our intuitions about the relationship between transparency and the permissibility of discrimination clash with moral theory, it is our intuitions which require refining, and not the theory. This is, at least, what I take Eidelson to imply. I would argue that while his responses are not entirely satisfactory, his account fares better when compared to alternative accounts. Indeed, while harm-based accounts face the dilemma of secretive and morally neutral racial

profiling, alternative views regarding the matter may be problematic all the same. For example, Mogensen notes that those who claim racial profiling to be intrinsically objectionable need to

> explain what makes racial profiling directed at black Americans or other similarly situated groups […] objectionable as it is generally regarded as being if not its characteristic harmful effects, given that we aren't strongly opposed to other forms of criminal profiling or statistical discrimination and, furthermore, accept that being black can in some cases be a contributing factor that leads to increased suspicion when police are searching for an individual matching a credible suspect description. (2019, 460.)

In other words, a proponent of such a view should be able to explain (i) what makes a sensitive trait such that it morally requires categorical blindness to it in decision-making while non-sensitive traits do not do so. This seems problematic given the fact that proxy discrimination in AD may be severely harmful, even though proxies do not fully map values of sensitive traits. Moreover, such a view should (ii) explain why categorical blindness is required in identical contexts and reasonable practices, such as affirmative action programs. Given the examination of common objections against sensitive characteristic -based statistical discrimination in chapter 3, it seems there is little room to argue that (i-ii) could be done by relying solely on notions about respect for individuality, unequal treatment, or the epistemic value of statistical evidence in relation to the sensitivity of a given trait.

However, Mogensen – due to what I take as a misinterpretation of Hellman's account – fails to see a third option. If one wishes to resolve the peculiar relationship between transparency and permissibility of profiling, it seems one could supplement one's view in some way that renders statistical discrimination on the basis of sensitive traits contingently objectionable while simultaneously preserving the notion that profiling is wrong even when conducted in secret. Hellman (2017) endorses this option and claims that demeaning acts – e.g., racial profiling – are morally objectionable even when carried out in secret because they express a derogatory *objective* meaning. Notably, Mogensen claims that Hellman's account would fail to identify secretive racial profiling as morally wrongful. Such an act, he argues, would express no demeaning message because no one is there to interpret the act. (Mogensen 2019, 459.) However, Mogensen fails to acknowledge Hellman's counterfactual notion according to which an act may be demeaning even in the absence of an interpreter: acts conducted by actors with sufficient social power have an objective meaning. Hellman explicitly addresses the issue of regarding the morality of secretive or uninterpreted acts in her work. She states that insofar "as the action, if known, would have this [derogatory] meaning, then it expresses denigration even when no [one] knows about it" (Hellman 2017, 105). Thus, her account indeed entails that secretive profiling is wrongful, contrary to Mogensen's claim. Yet because she

qualifies the claim by maintaining that only acts performed by actors with sufficient social power may express such a meaning, it is only contingently and not intrinsically so. That is, an identical act is not demeaning in every possible instance; only in those where it is conducted by someone with social power. Consequently, this would amount to a claim that unalloyed algorithmic discrimination, even if conducted in secret, would be morally objectionable, insofar as the discriminator possesses sufficient social power. For a proponent of the ET, this could prove a possible way out of the dilemma.

As noted in chapter 1.2.3, Hellman's account suffers from significant problems, however. The account fails to identify idiosyncratic instances of discrimination which involve socially non-salient groups – such as people with an uneven number of siblings – because they may not involve violations of respect-conventions. (Beeghly 2017, 90; Eidelson 2015, 88.) Arguably, there is something wrong with discriminating against people with three siblings as though they were categorically lesser in value in comparison to those with two siblings, however. Eidelson would claim that it fails to respect them *as* persons with equal standing, even in the absence of respect-conventions regarding the matter because it denies their equal value in comparison to others. Furthermore, the notion that acts may be demeaning even when conducted in secret seems to undermine the role of disrespect in Hellman's theory of discrimination. One could ask, how can an act demean if it goes uninterpreted by anyone and an act is demeaning in virtue of social convention(s)? This problematic notion is recognized by Lippert-Rasmussen, who states that Hellman's "position implies that an act may be impermissible in virtue of something – a free-floating, uninterpreted cultural meaning – that does not really affect people's lives" (2014, 131)[129]. On a related note, he also points out that Hellman's account exhibits a "problem of epiphenomenality" (Ibid., 136). The problem is that "while any wrongful act of discrimination is demeaning, some underlying factor explains both the fact that they are wrongful and that they are demeaning – the fact that discriminatory acts are demeaning does no explanatory work in relation to the wrongfulness of these acts" (Ibid.).

In light of these problems, Hellman's account seems to only function as a way of explaining the wrongness of a subset of all possible instances of wrongful discrimination. Eidelson's view, however, is more robust given its pluralistic nature. It identifies paradigmatic instances of discrimination and explains their wrongness by reference to a failure to either recognize the equal standing of persons or to afford sufficient normative weight for information regarding their individuality and autonomy in deliberation, and the subsequent actions taken on the basis thereof.

[129] One could naturally object against Eidelson's view on somewhat similar grounds in that his account rests on a universal moral principle; the moral demand of recognition respect which exists independently of interpretation and is rooted in people's equal moral standing as persons. However, I would argue that in Hellman's view this problem is more severe as her account is explicitly built on the notion that respect-conventions are socially grounded to begin with.

Some instances may even involve contempt, a knowing and ill-intended refusal to treat people as moral equals. Secondly, it does not face the problem of epiphenomenality. In the context of unalloyed statistical discrimination, the explanatory work is done by produced material and psychological *harms* and the (re)production of social patterns that manifest basic disrespect. Of course, one could argue that we should favor more straightforward and simplistic explanations in theory formation as opposed to a pluralist account – i.e., ones that reduce the moral wrongness of discrimination to a single kind of moral wrong, such as harm alone. However, discrimination is no straightforward phenomenon, as is perhaps exemplified by the topics addressed in this study. Demanding simplicity could be too strong of a requirement. Perhaps, then, one could embrace the idea that ethics of discrimination ought to consist of "a somewhat messy blend of deontological and consequentialist considerations", to quote Alexander (1992, 154).

On the basis of this examination, it seems that one has two plausible options with regard to transparency and the permissibility of profiling. One could concede that there are (at least hypothetical) morally neutral instances of secretive statistical discrimination against vulnerable groups; an option Eidelson prefers. Alternatively, one could maintain, as Hellman does, that acts may have objective derogatory meanings even when they are conducted in secret, rendering those acts morally objectionable. Weighing the problems of both accounts against each other, I have argued that Eidelson's account is preferable in this regard. It is not entirely satisfactory, however, and there is room for further philosophical inquiry into the peculiar relationship between harm-based accounts and transparency.

4.4. Tackling Algorithmic Discrimination: Why Means Matter

In this study, I have demonstrated that there is a plurality of ways in which wrongful algorithmic discrimination might arise, spanning from poor design choices to reliable and accurate modeling of systemic inequity. It seems, then, that the countermeasures taken against wrongful discrimination in AD should reflect this plurality. I will next provide a glimpse into what could be done.

Designers and decision-makers could be required to conduct careful risk-assessments and evaluations of the possibly harmful social impact of the used technology. These assessments and evaluations could be conducted before and during design processes and continued throughout the technology's lifespan. Problems residing within design processes and stemming from human-supervised use may require implementation of internal or external auditing mechanisms. (Kaminski 2019; Raji et al. 2020.) Documentation, regular reporting and extensive auditing could prevent possible intentional and unintentional discrimination taking place in design and development

processes. For example, auditing the data used by organizations could ensure that "the datasets used are not tainted [with human biases] prior to launching schemes based upon them, as the datasets will adversely impact minorities" (Zarsky 2014, 1393). Issues related to discrimination in design and human-in-the-loop use of ADSs may also require "de-biasing" human actors via cognitive training (Ibid., 1392) and ensuring that the teams in charge of technology design are diverse (cf. West et al. 2019), spanning across different demographics with their unique viewpoints regarding what constitutes 'fair' – or, as I would call it, *just* – technology.

ADSs could be causally modeled and designed to be inherently interpretable. Zarsky says that an interpretable model enables "the analyst to go beyond correlation and search for a theory that could uncover *causation*" (2013, 1520). This may aid in preventing wrongful discrimination as "interpretability and causation allow analysts to identify instances where the patterns used amount to illegal discrimination" (Ibid.) Designing algorithms in a manner that is specific and sensitive to the social context could be preferable, as "repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context" (Selbst et al. 2018, 4). Intersectional disadvantage in AD could be identified by testing models via using synthetically tailored data inputs, i.e., fake persons specified by conjunctions of vulnerable identity categories (cf. Raji et al. 2020, 9). In addition, as Chander (2016) noted, algorithmic affirmative action could be adopted as an explicit normative demand for AD in contexts where it does not conflict with legal norms and other ethical principles, such as public safety. This would amount to employing fairness metrics that explicitly aim to correct past wrongs that have led to inequalities of the present.

Legal solutions could be implemented to provide robust safeguards against discrimination. Wachter and Mittelstadt (2019) have argued for the incorporation of a novel data protection right, "a right to reasonable inferences", which would require data controllers to provide a threefold, *ex-ante* justification for using inferred data in AD. Those in control of data would be required to justify not only "why certain data form a normatively acceptable basis from which to draw inferences" but also "why these inferences are relevant and normatively acceptable" for use in a given context (Ibid., 2). Furthermore, they should ensure that the inferential methods (i.e., statistical and ML methods) "are accurate and statistically reliable" (Ibid.). Wachter (forthcoming) has also noted that individuals may be discriminated against due to their *presumed* similarity or affinity to other individuals that belong to protected groups (e.g. one's similarity to a cluster of individuals of some ethnicity). Providing legal safeguards against this type of discrimination-by-association "would help overcome the argument that inferring one's 'affinity for' and 'membership in' a protected group are strictly unrelated" (forthcoming, 2). Arguably, this solution would align the legal concept of

discrimination with the philosophical notion that discrimination as an act is *actual-properties independent*: it is not necessary that individuals belong to some group in order to be discriminated against in virtue of the discriminator's presumption that this is the case. Notably, Wachter's proposed solution would also prevent cases where individuals have to "'out' themselves" as members of vulnerable groups (e.g. LGBTIQA+) to claim their legal protection against discrimination (Ibid., 2), thus ensuring that they may keep their identity private and remain safe from public prejudice.

I have suggested that the belief-dependent harms of algorithmic discrimination are closely connected to what a given practice supported by AD *expresses* in relation to context and the respect-conventions that govern possibilities for permissible conduct in that context. In this sense, it is the social meaning of a given practice (in addition to other harms) that may in some cases render that practice wrongful as opposed to whether sensitive information enters the decision-making process, either explicitly or via proxy. This has an interesting implication for the moral evaluation of algorithmic discrimination: the possible wrongness of using a given "biased" ADS will also depend in part on *what that ADS is used for*. In a sense, the social meaning of a practice or policy may partly determine whether it is "biased" in the moral sense. For example, assume that predictive policing algorithms would accurately estimate that minority members are in higher risk for committing crimes, or that neighborhoods primarily occupied by them are places where crime might occur. Rather than using such algorithms to engage in more granular targeting of potential suspects and to find areas where "stop and frisk" tactics might lead to discovering criminal activity, the same algorithms could be used for other preventive and supportive methods which promote well-being and human flourishing. Andrew Guthrie Ferguson has endorsed this idea, arguing that predictive policing algorithms could be used to "map out the social and economic vulnerabilities in an area" and employ them as a way of remedying those vulnerabilities by other means than policing (2016, 1184). As "certain individuals face external challenges that increase the chance for violence", the same algorithms could be used "to identify people in need of other social services (education, employment, mental health services)" (Ibid.). He acknowledges that while such uses of ADSs could also have stigmatizing effects – e.g. they could reinforce certain similar stereotypes as policing methods do – this could at least have a better overall effect than the punitive methods that are currently used (Ibid.).

Ferguson hints towards an important notion. The measured "fairness" of an algorithm may not in all cases make as significant a moral difference as opposed to what the practice informed by that algorithm expresses. The same statistical model that accurately maps out correlations between crime and race can be interpreted either as a justification for surveillance and possible detainment, or as a call for supportive policy and redistributive justice; algorithmic affirmative action, as Chander (2016) called it. The latter, I would argue, is better aligned with the demand that we show respect for

individuals' autonomy, equality, and well-being. Hence, it is not only the algorithms and models, but also the *means* that matter for achieving a given praiseworthy goal.

4.5. Chapter Summary

In this chapter, I have considered Eidelson's Broad Harms Argument as a way of explaining why unalloyed algorithmic discrimination is morally wrong when it has a disparate impact on some vulnerable group. According to this view, use of algorithms that has a disparate impact on groups specified by sensitive traits – even if well-intended and carefully conducted in order to achieve a reasonable aim, such as preventing crime – will be wrong if and when it results in significant material and psychological harms, and when it sustains or further cements social conditions that encourage behavior that is disrespectful in the basic sense. This account allows that the relevant psychological harms of AD do not necessarily concern only those who are disadvantaged, but also those who feel demeaned, insecure or insulted as a result of such practices being conducted. These harms may be expressive in the sense that they are a result of violations of respect-conventions. For example, while inferences from race to crime in predictive policing and recidivism risk assessment may be epistemically justified given past data, relying on such inferences is conventionally disrespectful due to a long history racially biased policing and institutional racism in the justice system. As such, it is derogatory and demeans people of color. AD may also exacerbate stigma related to vulnerable groups in that, when modeled into ADSs, wrongs of the past become a seemingly rational yet deceptive "ground truth" for future decisions. That is, algorithms may portray vulnerable groups as fundamentally less deserving, dangerous and unworthy, even though what the algorithms have learned is, in effect, a representation of how members of these groups have been (wrongfully) treated in the past.

The presented account is robust in virtue of leaving open the possibility that some instances of algorithmic discrimination may be morally benign. For example, algorithmically conducted practices that aim to improve the well-being of vulnerable groups and ones that stem from legitimate business interests and are not accompanied by significant harms may not be wrongful even if they involve sensitive classification or the algorithms are biased in the statistical sense. This also explains why algorithmic discrimination against non-salient, synthetic groups is not often morally wrong. I would argue that this is a desirable implication of the account given that many common and reasonable practices, such as focusing preventive mental health services to those in need through statistical screening or calculating individualized insurance premiums, rely on statistical discrimination. I noted that the account entails a dilemma under certain circumstances: given that

expressive harms are belief-dependent, secretive algorithmic profiling may not be morally wrong. I argued that this is a significant problem for the account, but insofar as other alternatives go, one might have to bite the bullet in this regard. Further research is needed to provide a fully satisfactory account of the ethics of secretive algorithmic profiling. Lastly, I offered a brief overview of methods for tackling algorithmic discrimination. Prominent methods include careful auditing and oversight, promoting diversity in technology-design, explicit adoption of fairness measures that serve a function of redistributive justice, as well as legal measures that provide safeguards against discrimination in the digital era. Given that the harms of algorithmic discrimination are dependent on the context and social meaning of AD practices, I also suggested that *how* we use algorithms in decision-making is also relevant. To establish practices that support individuals' autonomy and well-being, it may be required that we revise approaches to, say, crime prevention; instead of using algorithms as means for surveillance and punishment, they could inform practices that aim to abolish the underlying social causes of crime.

5. Conclusions

In this study, I have analyzed the phenomenon of algorithmic discrimination and ethics thereof. The literature on AI ethics shows that there are several ways in which disparate impact may follow from the use of "biased" algorithms in decision-making. The identification of these distinct mechanisms of algorithmic discrimination have been accompanied by claims as to why such discrimination is wrong. However, these claims have inherited questions and conceptual problems that have long characterized the discussion on discrimination in fields of legal and moral theory. To clear some conceptual issues exemplified by the discourse on algorithmic discrimination, I approached the subject from the point of view of moral philosophy. Specifically, I applied a slightly modified version of Benjamin Eidelson's disrespect-based theory of discrimination to provide an account of (i) what constitutes discrimination in algorithmic decision-making, and (ii) how one could explain the wrongness of distinct instances of algorithmic discrimination. According to Eidelson's account, intrinsically wrongful discrimination involves either a failure or an upright refusal to respect the personhood of those discriminated against. A discriminator will in such cases disrespect the equal moral standing of some individual or disrespect her standing as an autonomous individual. Contingently wrongful discrimination does not involve the aforementioned qualities but produces significant harm. It will be wrong even if these harms are unpredictable to the otherwise careful and reasonable discriminator. This theory, I suggested, offers a defensible account of the wrongness of algorithmic discrimination, albeit it has its problems.

Exploring different pathways or mechanisms of algorithmic discrimination, I noted that decisions made with respect to what is being predicted and how, what data and predictive features are to be used, and what machine learning methods to employ, among other things, may either exacerbate or mitigate possible disparate impact resulting from algorithmic decision-making. Algorithmic bias that leads to disparate impact on some group may follow from deliberate or unintentional design. Human cognitive biases, alongside problems with trust, psychological priming and interpretation, might exacerbate discriminatory outcomes when algorithms are put to use. However, as algorithmic decision-making involves differential treatment informed by statistical evidence *by definition*, it also involves statistical discrimination whenever some group is comparatively disadvantaged as a result. Thus, there is no single phenomenon or discriminatory process that the concept of algorithmic discrimination captures, but many.

Analyzing the distinct mechanisms that may result in disparate impact, I noted, that a distinction between statistical and moral bias is required in order to understand different forms of algorithmic discrimination. Algorithms may exhibit *statistical model bias* – i.e., deviations between

population values and the statistical information captured in the model – but also *moral bias*. The latter refers to a deviation between the models performance and some moral principle: an algorithm may be accurate yet still reproduce existing inequality that has been wrought by historical discriminatory practices, such as racially biased policing, or which is a symptom of underlying unjust social conditions, such as poverty. Because different types of bias may interact and counteract in the design and use of algorithmic decision-making systems, I argued that the moral evaluation of an instance of algorithmic discrimination requires a fine-grained analysis of these interactions. Employing the distinction between *tainted* and *unalloyed* statistical discrimination, I argued that the wrongness of some instances of algorithmic discrimination is traceable to distinct acts that are conducted in different dimensions in both the design process of algorithmic decision-making systems and the "human-in-the-loop" use thereof. They should be understood as wrongful instances of algorithmic discrimination in that they are tainted by disrespect towards some group, and not explained by a piece of statistical evidence as such. Some of these instances may be the result of second-order discrimination, both intentional and malicious, as well as unintentional albeit still wrongful, in the design process. Instances of "human-in-the-loop" algorithmic decision-making constitute direct discrimination when a decision is not exhaustively explained by the accessed statistical evidence but, rather, biased human action or adoption of AD as a means for conducting blatantly oppressive policies. Accordingly, the *bad actor view* (BAV) and the *bad technology view* (BTV), as I called them, identify instances of algorithmic discrimination that may be wrongful under specific conditions. Explaining which instances are wrong and which are benign cannot be done only by examining whether algorithmic decision-making involves sensitive or proxy classification, or whether disparate impact is a deliberate effort as opposed to an unintended result of it, however. As the *structural discrimination view* (SDV) maintained, disparate impact may follow from even careful modeling conducted in good intent when data mining leads to a discovery of a proxy for sensitive group-membership, which then predicts a certain outcome. Such instances, if unaccompanied by contingent moral issues, can be understood as unalloyed instances of algorithmic discrimination. Tainted discrimination is often more straightforwardly wrong, as it violates an individual's right to be treated as a moral equal or involves the use of inaccurate data, for example. The wrongness of unalloyed instances of algorithmic profiling, however, cannot be exhaustively explained by reference to unequal treatment in design, carelessness, sensitive classification or issues with inaccuracy as these are issues contingently associated with acts falling under the category of statistical discrimination.

As I noted, the notion that not all instances of wrongful algorithmic discrimination can be explained by malicious discriminatory intent and poor design leaves room for the "Nothing Personal" Argument: insofar as reliable methods and accurate data have been used, and the practice

has been respectfully conducted, some seemingly problematic instances of algorithmic decision-making could be morally neutral despite the fact that they may have a disparate impact on certain groups. The possibility of unalloyed algorithmic discrimination also raises a second question, namely, the Synthetic Groups Question (the SGQ): if some group is always comparatively disadvantaged as a result of algorithmic decisions (i.e., the individuals that constitute the class of comparatively worse decisions), why are not *all* instances of algorithmic discrimination wrong?

I considered two types of principled arguments that would allegedly explain why unalloyed discrimination is wrong and, consequently, answer the SGQ. The Inaccuracy Argument stated that algorithmic discrimination is wrong because people are treated on the basis of inaccurate profiles. This does not exhaustively explain the wrongness of statistical discrimination, however, because statistical generalizations may in fact be highly accurate. The Causal Connection Argument claimed that algorithmic discrimination involves epistemically unwarranted inferences, such as inferences from race to crime, because statistical evidence tells us nothing about causality – such inferences, allegedly, do not constitute individualized evidence. I argued that grounding the distinction between individualized and non-individualized evidence in causality (or lack thereof) leads to problematic implications. Assuming reasonable accuracy, statistical evidence does epistemically warrant differential treatment. Furthermore, the argument does not explain why there would be an *intrinsic* moral difference when those inferences involve sensitive traits. The Objectification Argument stated that algorithmic profiling leads to treating people as nothing but artefacts of their group-membership. I argued that as the data required to capture individuals' character may be inaccessible to data miners, algorithmic discrimination may not involve disrespect for individuality. Furthermore, as disrespect for individuality is a non-comparative notion while wrongful discrimination is comparative, the set of people whose individuality is not afforded adequate respect and those who are discriminated against may not be identical. A second type of principled objections maintained that wrongful algorithmic discrimination is inherently unfair in that it involves unequal treatment or partiality. I noted that some objections from unfairness are in effect demands for redistributive justice. This relates to how Eidelson understands some prohibitions of 'indirect' discrimination: these legal prohibitions do not *only* aim to prevent wrongful instances of discrimination *per se*, but also serve to improve some group's status. This may be a morally justified cause but confuses objections from unfairness with claims regarding the harm algorithmic discrimination does or the social conditions it reproduces. I also explored the more philosophical yet common idea underlying fairness considerations, namely, that it is wrong to treat people differently on the basis of traits that they are not responsible for or cannot alter. I argued that these objections do not hold under closer scrutiny: One could be responsible for a trait and still be treated wrongly,

because responsibility for a trait and responsibility for a discriminatory act are two different objects of responsibility. Insofar as mutability goes, it does not ground a principled objection against discrimination as it makes a morally significant difference only to the extent that profiling practices are epistemically transparent for an individual (which they often are not given that their practical value hinges on opacity). Moreover, there are multiple discriminatory policies that we are generally willing to accept, even though they involve differential treatment on the basis traits that we are not responsible for or cannot change.

I proceeded to consider Eidelson's Broad Harms Argument as a way of defeating the "Nothing Personal" Argument and answering the Synthetic Groups Question. The account states that a given practice of statistical discrimination will be wrong if the material and psychological harms it produces, and the undesirable broad social consequences it has, outweigh its benefits. If one adopts Eidelson's view, algorithmic discrimination can be understood as only contingently wrongful and not intrinsically so. I suggested that some paradigmatic examples of unalloyed algorithmic discrimination can be considered wrongful in light of this argument. Considering predictive policing, recidivism risk assessment and credit-scoring in particular, I argued that the material harms and the feelings of shame, denigration and insecurity they result in, as well as the stigma and patterns of disrespectful social conduct they sustain and possibly amplify, may render them wrongful in light of a careful qualitative analysis regarding the experiences of harm that accompany them. However, emergent discrimination against synthetic groups may be morally neutral or even praiseworthy according to this account. This is because it does not result in similar broad harms as those instances which bear a loaded social meaning in virtue of background injustice suffered by the groups that are discriminated against. Thus, the Broad Harms Argument can both defeat the "Nothing Personal" Argument and answer the Synthetic Groups Question.

I noted that lack of transparency creates a problem for the account explored in this study in that one has to concede that secretive algorithmic profiling which has a disparate impact on some group may be morally neutral or even praiseworthy. (This dilemma is pertinent only if the material harms do not outweigh the benefits of profiling, however.) While this conclusion is perhaps even more problematic when considered in the context of algorithmic discrimination, the account seems to nevertheless constitute a more robust alternative when compared to other theories. As such, biting the bullet with respect to the dilemma of secretive algorithmic profiling may be preferable from a theoretical point of view. The dilemma shows that further research is required into the phenomenon of algorithmic discrimination and ethics thereof.

To conclude the study on a hopeful note, I considered some ways of tackling algorithmic discrimination at various stages of design and use of algorithmic decision-making systems. Ranging

from audits and intersectional fairness evaluations to legal solutions, these countermeasures could be implemented and further developed in order to prevent wrongful discrimination in the design and use of algorithmic technologies. Drawing on the findings of this study, I concluded that even "biased" algorithms could be used for practices that support – rather than infringe – human dignity and autonomy. Instead of using data that reflects background injustice as a justification for the reproduction of systemic inequality and for pushing individuals in positions of vulnerability deeper into the margins, the same data could be used to inform practices that aim to correct wrongs wrought by past discrimination and which serve to establish social conditions that promote equality between individuals and groups.

Bibliography

Alexander, L. (1992). What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review*, 141(1), pp. 149-219.

Alpaydin, E. (2016). *Machine Learning*. Cambridge (Mass.): MIT Press.

Altman, A. (2016). Discrimination. *The Stanford Encyclopedia of Philosophy*, Zalta E. N. (ed.). Retrieved from https://plato.stanford.edu/archives/win2016/entries/discrimination/. [Accessed 9.12.2019]

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accessed 28.1.2020]

Arkoudas, K. & Bringsjord, S. (2014). Philosophical foundations. *The Cambridge Handbook of Artificial Intelligence*, Frankish, K. & Ramsey, W. M. (ed.). Cambridge: Cambridge University Press.

Barocas, S. (2014). Data mining and the discourse on discrimination. *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining 2014*, pp. 1–4.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, pp. 671–732.

Beeghly, E. (2017). Discrimination and Disrespect. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

(2018). Failing to Treat Persons as Individuals. *Ergo, an Open Access Journal of Philosophy*, 5, pp. 687–712.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *arXiv preprint*, arXiv:1712.03586.

Boden, M. A. (2014). GOFAI. *The Cambridge Handbook of Artificial Intelligence*, Frankish, K. & Ramsey, W. M. (ed.). Cambridge: Cambridge University Press.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, pp. 77–91.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), pp. 1–12.

Chan, R. (2019, July 19). Here's what you need to know about Palantir, the secretive $20 billion

dollar data-analysis company whose work with ICE is dragging Amazon into controversy. *Business Insider*. Retrieved from https://www.businessinsider.com/palantir-ice-explainer-data-startup-2019-7?r=US&IR=T. [Accessed 18.2.2020.]

Chander, A. (2016). The racist algorithm. *Michigan Law Review*, 115(6), pp. 1023–1045.

Citron, D. K. & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89, pp. 1–33.

Coeckelbergh, M. (2019). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and engineering ethics*, pp. 1-18.

Collins, H. (2017). Discrimination and the Private Sphere. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K (ed.). New York: Routledge.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*, arXiv:1808.00023.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Conference on Fairness, Accountability and Transparency*, pp. 134–148.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review,* 43(6), pp.1241–1299.

Dallas, C. (2018, March 25). COMPAS, revisited. *Medium*. Retrieved from https://medium.com/@dallascard/compas-revisited-f35464686fe2. [Accessed 15.2.2020]

Danks, D. (2014). Learning. *The Cambridge Handbook of Artificial Intelligence*, Frankish, K. & Ramsey, W. M. (ed.). Cambridge: Cambridge University Press.

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *IJCAI*, pp. 4691–4697.

Darwall, S. L. (1977). Two kinds of respect. *Ethics*, 88(1), pp. 36–49.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, business news. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. [Accessed 23.9.2019]

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*. Retrieved from https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/. [Accessed 28.1.2020]

Dobbe, R. & Whittaker, M. (2019, October 17). AI and Climate Change: How they're connected, and

what we can do about it. *AI Now Institute*. Retrieved from https://medium.com/@AINowInstitute/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c. [Accessed 18.2.2020.]

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances, 4*(1).

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, 58(6), pp. 697–718.

Eidelson, B. (2013). Treating People as Individuals. *Philosophical Foundations of Discrimination Law*, Hellman, D. & Moreau, S. (ed.). Oxford: University Press.

(2015). *Discrimination and Disrespect*. Oxford: University Press.

(Forthcoming). Respect, Individualism, and Colorblindness. *Yale Law Journal*, forthcoming. Retrieved from https://ssrn.com/abstract=3473832. [Accessed 5.12.2019.]

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

European Union. *Charter of Fundamental Rights of the European Union*. (2012, October 26, 2012/C 326/02). Retrieved from https://www.refworld.org/docid/3ae6b3b70.html. [Accessed 15.11.2019]

Ferguson, A. G. (2016). Policing predictive policing. *Washington University Law Review*, 94, pp. 1109–1189.

Flew, A. (1990). Three Concepts of Racism. *Encounter*, 75, pp. 63–66.

Franklin, S. (2014). History, motivations, and core themes. *The Cambridge Handbook of Artificial Intelligence*, Frankish, K. & Ramsey, W. M. (ed.). Cambridge: Cambridge University Press.

Fricker, M. (2003). Epistemic justice and a role for virtue in the politics of knowing. *Metaphilosophy*, 34(1-2), pp. 154–173.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv preprint*, arXiv:1609.07236.

Garcia, J. L. A. (2017). *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), pp. 121–127.

Gosepath, S. (2011). Equality. *The Stanford Encyclopedia of Philosophy*, Zalta E. N. (ed.). Retrieved from https://plato.stanford.edu/archives/spr2011/entries/equality/. [Accessed 20.2.2020]

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), pp. 1–24.

Han, H. & Jain, A. K. (2014). Age, Gender and Race Estimation from Unconstrained Face Images. *MSU Technical Report*. [Accessed 23.8.2019].

Enoch, D. & Fisher, T. (2015). Sense and sensitivity: Epistemic and instrumental approaches to statistical evidence. *Stanford Law Review*, 67, pp. 557–611.

Ethics Guidelines for Trustworthy AI. (2019). Independent High-Level Expert Group on Artificial Intelligence set up by the European Comission. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Halldenius, L. (2017). Discrimination and irrelevance. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Hellman, D. (2008). *When is Discrimination Wrong?* Oxford: Oxford University Press.

(2017). Discrimination and Social Meaning. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7), pp. 900–915.

Jezzard, P., Matthews, P. M., & Smith, S. M. (Eds.). (2001). *Functional MRI: an introduction to methods* (Vol. 61). Oxford: Oxford University Press.

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, pp. 389–399.

Kaminski, M. E. (2019). Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability. *Southern California Law Review*, 92(6), pp. 1529–1616.

Kant, I., (1785). *Grundlegung zur Metaphysik der Sitten*, translated as "Groundwork of the Metaphysics of Morals". *Immanuel Kant Practical Philosophy* (1996), Gregor, M. (trans. and ed.), New York: Cambridge University Press.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2017). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint*, arXiv:1711.05144.

Kelleher, J. D. & Tierney, B. (2018). *Data Science*. Cambridge: MIT Press.

Kelly, D. & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, *3*(3), pp. 522-540.

Khaitan, T. (2017). Indirect Discrimination. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, pp. 656–666.

Kingma, D. P., Mohamed, S., Rezende, D. J. & Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in neural information processing systems 27*, pp. 3581–3589.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint*, arXiv:1609.05807.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). *Counterfactual fairness. Advances in Neural Information Processing Systems*, pp. 4066–4076.

Langton, R., Haslanger, S. & Anderson, L. (2012). Language and Race. *The Routledge Companion to the Philosophy of Language*, Russell G. & Fara, D. G. (eds.), pp. 753–767. New York: Routledge.

Lazenby, H. & Butterfield, P. (2017). Discrimination and the Personal Sphere. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), pp. 436–444.

Leslie, S-J. (2012). Generics. *The Routledge Companion to the Philosophy of Language*, Russell G. & Fara, D. G. (eds.), pp. 355–365. New York: Routledge.

Lever, A. (2007). What's wrong with racial profiling? Another look at the problem. *Criminal Justice Ethics*, 26(1), pp. 20–28.

Lippert-Rasmussen, K. (2006). The badness of discrimination. *Ethical Theory and Moral Practice*, *9*(2), pp. 167-185.

(2007). Nothing personal: On statistical discrimination. *Journal of Political Philosophy*, 15(4), pp. 385-403.

(2011). ”We are all Different”: Statistical Discrimination and the Right to be Treated as an Individual. *The Journal of ethics*, 15(1-2), pp. 47–59.

(2014). *Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination*. Oxford University Press.

(2019). Respect and discrimination. *Moral Puzzles and Legal Perplexities*, Hurd, H. M. (ed.). Cambridge: University Press.

Lockwood, B. (2020, January 30). The History of Redlining. *ThoughtCo*. Retrieved from https://www.thoughtco.com/redlining-definition-4157858. [Accessed 3.3.2020]

Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). Causal reasoning for algorithmic fairness. *arXiv preprint*, arXiv:1805.05859.

Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, 13(5), pp. 14-19.

Malnick, E. (2019, December 28). Police using AI the with ‘troubling’ secrecy, says former

MI5 chief. *The Telegraph*. Retrieved from https://www.telegraph.co.uk/politics/2019/12/28/police-using-ai-tech-troubling secrecy-says-former-mi5-chief/. [Accessed 30.12.2019]

Mann, M. & Matzner, T. (2019). Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society*. Retrieved from https://doi.org/10.1177/2053951719895805.

McCulloch, W. & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5, pp. 115–133.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint*, arXiv:1811.07867.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2).

Mogensen, A. (2019). Racial Profiling and Cumulative Injustice. *Philosophy and Phenomenological Research*, 98(2), pp. 452–477.

Moles, A. (2017). Discrimination and desert. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Moore, M. (2019). Causation in the Law. *The Stanford Encyclopedia of Philosophy*, Zalta E. N. (ed.). Retrieved from https://plato.stanford.edu/archives/win2019/entries/causation-law/. [Accessed 9.10.2019]

Moreau, S. (2010). What is discrimination?. *Philosophy & Public Affairs*, 38(2), pp. 143-179.

(2017). Discrimination and freedom. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Mozur, P. (2019, April 14). One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. *The New York Times*. Retrieved from https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html. [Accessed 3.3.2020]

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2), pp. 175–220.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: New York University Press.

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 13.

Packin, N. G. (2019). Algorithmic Decision-Making: The Death of Second Opinions?. *New York University Journal of Legislation and Public Policy* (Forthcoming).

Pincus, F. L. (1996). Discrimination comes in many forms: Individual, institutional, and structural. *American Behavioral Scientist*, 40(2), pp. 186–194.

Puddifoot, K. (2017). Epistemic Discrimination. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Raji, I. D., Smart A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. & Parker, B. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Conference on Fairness, Accountability, and Transparency (FAT* '20*).

Risse, M., & Zeckhauser, R. (2004). Racial profiling. *Philosophy & Public Affairs*, 32(2), pp. 131–170.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206–215.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge (Mass.): Harvard University Press.

Schauer, F. (2006). *Profiles, Probabilities, and Stereotypes*. Cambridge (Mass.): Harvard University Press.

(2017). Statistical (and Non-Statistical) Discrimination. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Selbst, A. D. (2017). Disparate impact in big data policing. *Georgia Law Review*, 52, pp. 109–195.

Selbst, A. D., Boyd, D., Friedler, S., Venkatasubramanian, S. & Vertesi, J. (2018). Fairness and Abstraction in Sociotechnical Systems. *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 1(1), Forthcoming. Retrieved from https://ssrn.com/abstract=3265913.

Sen, A. (1979). Equality of what?. *The Tanner lecture on human values*, 1, pp. 197–220.

Silva, S. & Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon (1960-)*, 55(1-2), pp. 9–37.

Springer, A., Garcia-Gathright, J., & Cramer, H. (2018). Assessing and Addressing Algorithmic Bias – But Before We Get There. *2018 AAAI Spring Symposium Series*.

Stanford, S. (2018, October 2; updated 2019, May 15). The Best Public Datasets for Machine Learning and Data Science. *Medium*. Retrieved from https://medium.com/towards-artificial-intelligence/the-50-best-public-datasets-for-machine-learning-d80e9f030279. [Accessed 30.8.2019]

Stoljar, N. (2017). Discrimination and Intersectionality. *The Routledge Handbook of the Ethics of Discrimination*, Lippert-Rasmussen, K. (ed.). New York: Routledge.

Stop LAPD Spying Coalition. (2013). "to observe and to suspect" – A People's Audit of the Los Angeles Police Department's Special Order 1. Retrieved from https://stoplapdspying.org/wp-content/uploads/2013/04/PEOPLES-AUDIT-UPDATED-APRIL-2-2013-A.pdf. [Accessed 8.1.2020]

Sun, R. (2014). Connectionism and neural networks. *The Cambridge Handbook of Artificial Intelligence*, Frankish, K. & Ramsey, W. M. (ed.). Cambridge: Cambridge University Press.

Talbott, W. (2016). Bayesian Epistemology. *The Stanford Encyclopedia of Philosophy,* Zalta, E. N. (ed.). Retrieved from https://plato.stanford.edu/archives/win2016/entries/epistemology-bayesian/>. [Accessed 9.12.2019]

Thomsen, F. K. (2013). But Some Groups Are More Equal Than Others – A Critical Review of the Group Criterion in the Concept of Discrimination. *Social Theory and Practice*, 39(1), pp. 120–146.

Verma, S. & Rubin, J. (2018). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare),* pp. 1–7.

Väyrynen, P. (2019). Thick Ethical Concepts. *The Stanford Encyclopedia of Philosophy*, Zalta, E. N. (ed.). Retrieved from https://plato.stanford.edu/archives/sum2019/entries/thick-ethical-concepts/. [Accessed 20.8.2019.]

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GPDR. *Harvard Journal of Law & Technology*, 31, pp. 841–887.

Wachter, S. & Mittelstadt, B. (2019). A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review,* 2019(2). Retrieved from https://ssrn.com/abstract=3248829.

Wachter, S. (forthcoming). Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *Berkeley Technology Law Journal*, 35(2), 2020 (Forthcoming). Retrieved from https://ssrn.com/abstract=3388639.

West, S.M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. *AI Now Institute*. Retrieved from https://ainowinstitute.org/discriminatingsystems.html.

Young, I. (1990). *Justice and the Politics of Difference*. Princeton: Princeton University Press.

Zarsky, T. Z. (2013). Transparent predictions. *University of Illinois Law Review*, 4, pp. 1503–1570.

Zarsky, T. Z. (2014). Understanding discrimination in the scored society. *Washington Law Review*, 89, pp. 1375–1412.